Some Recent Developments on Nonparametric Econometrics*†

Zongwu Cai
Department of Mathematics and Statistics and Department of Economics,
University of North Carolina at Charlotte, Charlotte, NC 28223, USA
Wang Yanan Institute for Studies in Economics, Xiamen University, China
Qi Li
Department of Economics, Texas A&M University
College Station, Texas 77843-4228, USA

In this paper we survey some recent developments of nonparametric econometrics in the following areas: (i) Nonparametric estimation of regression models with mixed discrete and continuous data; (ii) Nonparametric models with nonstationary data; (iii) Nonparametric models with instrumental variables; (iv) Nonparametric estimation of conditional quantile functions. In each of the above areas we also point out some open research problems.

Forthcoming in *Advances in Econometrics*

*We thank the referees for their careful reading of the manuscript and for their helpful comments. We also thank the Econometric Seminar audiences at the University of Guelph, and the participants in the Advances in Econometrics conference for helpful comments on this paper.

†Cai's research was supported, in part, by the National Science Foundation grant DMS-0404954 and the National Nature Science Foundation of China grant #70871003, and funds provided by the University of North Carolina at Charlotte, the Cheung Kong Scholarship from Chinese Ministry of Education, the Minjiang Scholarship from Fujian Province, China and Xiamen University. Li's research is partially supported by the National Nature Science Foundation of China grant #70773005.

1 Introduction

There is a growing literature in nonparametric econometrics in the recent two decades. Given the space limitation it is impossible to survey all the important recent developments in nonparametric econometrics. Therefore, we choose to limit our focus on the following areas. In Section 2 we review the recent developments of nonparametric estimation and testing of regression functions with mixed discrete and continuous covariates. We discuss nonparametric estimation and testing of econometric models for nonstationary data in Section 3. Section 4 is devoted to surveying the literature of nonparametric instrumental variable models. We review nonparametric estimation of quantile regression models in Section 5. In Sections 2 to 5 we also point out some open research problems, which might be useful for graduate students to review the important research papers in this filed and to search for their own research interests, particularly dissertation topics for doctoral students. Finally, in Section 6 we highlight some important research areas that are not covered in this paper due to space limitation. We plan to write a separate survey paper to discuss some of the omitted topics.

2 Models With Discrete And Continuous Covariates

In this section, we mainly focus on analysis of nonparametric regression models with discrete and continuous data. We first discuss estimation of a nonparametric regression model with mixed discrete and continuous regressors, and then we focus on a consistent test for parametric regression functional forms against nonparametric alternatives.

2.1 Nonparametric Regression Models With Discrete And Continuous Covariates

We are interested in estimating the following nonparametric regression model

$$Y_i = g(X_i) + u_i, \qquad (i = 1, ..., n) \tag{2.1}$$

where $X_i = (X_i^c, X_i^d)$, $X_i^c \in \Re^q$ is a continuous random variable of dimension q ($q \geq 1$), and X_i^d is a discrete random variable of dimension r ($r \geq 0$). We will only consider independent and identically distributed data case in Section 2. Let X_{is}^d denote the s-th component of X_i^d. We consider two possibilities: X_{is}^d can be an ordered and un-ordered discrete variable. If X_{is}^d is un-ordered, $X_{is}^d \in \mathcal{D}_s = \{a_1, a_2, ..., a_{c_s}\}$ with c_s taking distinct different values and $c_s \in \mathcal{N}$, where $\mathcal{N}$ denotes the set of positive integers. Here we allow for the possibility that $c_s = \infty$. If $c_s = \infty$, we need to add a condition that $\inf_{x_s^d \neq x_{s'}^d; x_x^d, x_{s'}^d \in \mathcal{D}_s} |x_s^d - x_{s'}^d| \geq \delta > 0$ so that x_s^d can take at most countably infinitely many different values, and there is only finite many distinct points of x_s^d in any bounded interval.

The conventional approach dealing with the discrete variable is to split the sample into many parts sorted by different discrete cells. Then one uses the data falling into a given discrete cell to estimate the conditional mean function of Y given the remaining continuous variables. However, this sample splitting method may give unreliable estimation results or even become infeasible when the number of discrete cells is not small compared with the sample size. In a seminal paper, Aitchison and Aitken (1976) proposed a novel method of smoothing discrete variables in estimating a discrete probability function. Hall, Racine and Li (2004), Racine and Li (2004), Hall, Li and Racine (2007) generalized Aitchison and Aitken's smoothing method to the problem of estimating a conditional density function or a conditional mean function. Their proposed smoothing method avoids the sample splitting problem and therefore remains a feasible estimation method when the number of discrete cells is comparable or even larger than the sample size. An additional advantage of smoothing the discrete variables is that, as shown by Hall, Racine and Li (2004), and Hall, Li and Racine (2007), irrelevant covariates can be automatically smoothed out (i.e., removed) from a conditional density or a regression model.

We now introduce the kernel smoothing function for discrete variables. The kernel function associated with un-ordered discrete variable X_{is}^d is given by $l(X_{is}^d, x_s^d, \lambda_s) = 1$ if $X_{is}^d = x_s^d$; and $l(X_{is}^d, x_s^d, \lambda_s) = \lambda_s$ if $X_{is}^d \neq x_s^d$, where λ_s is the smoothing parameter. If X_{is}^d is an ordered discrete variable, we use the following kernel function: $l(X_{is}^d, x_s^d, \lambda_s) = \lambda_s^{|X_{is}^d - x_s^d|}$. Whether x_s^d is either ordered or un-ordered, when $\lambda_s = 0$, the kernel function becomes an indicator function, i.e., $l(X_{is}^d, x_s^d, 0) = \mathbf{1}(X_{is}^d = x_s^d)$, where $\mathbf{1}(A)$ denotes an indicator function that takes value one if event A holds true, and zero otherwise. Also, when $\lambda_s = 1$, $l(X_{is}^d, x_s^d, 1) \equiv 1$ is a constant function. The range of λ_s is $[0, 1]$ for all $s = 1, ..., r$. The product kernel for the discrete variables X^d is $L(X_i^d, x^d, \lambda) = \prod_{s=1}^r l(X_{is}^d, x_s^d, \lambda_s)$. For the continuous variable $X^c = (X_1^c, \ldots, X_q^c)$, we use the product kernel given by $W_h(x^c, X_i^c) = \prod_{s=1}^q h_s^{-1} w\left((x_s^c - X_{is}^c)/h_s\right)$, where $w(\cdot)$ is a symmetric and univariate density function, and $0 < h_s < \infty$ is the smoothing parameter for x_s^c.

The kernel function for the mixed regressor case $X = (X^c, X^d)$ is simply the product of W and L, i.e., $K(x, X_i) = W_h(x^c, X_i^c) L(x^d, X_i^d, \lambda)$. Thus we estimate $g(x) = E(Y|X = x)$ by the Nadaraya-Watson (local constant) method, defined as

$$\widehat{g}(x) = \frac{\sum_{i=1}^n Y_i K(x, X_i)}{\sum_{i=1}^n K(x, X_i)}. \tag{2.2}$$

It is easy to see that if $\lambda_s = 0$ for all $s = 1, ..., r$, then the discrete kernel function becomes an indicator function, i.e., $L(X_i^d, x^d, 1) = \mathbf{1}(X_i^d = x^d)$. $\widehat{g}(x)$ defined in (2.2) reduces to the conventional frequency estimator of $g(x)$. Also, if $\lambda_s = 1$ for some $s \in \{1, ..., r\}$, since $l(X_{is}^d, x_s^d, 1) \equiv 1$, in this case $\widehat{g}(x)$ becomes unrelated to x_s^d, i.e., the covariate x_s^d is completely removed from the regression model. Similarly, for the continuous variable x_s^c, if h_s is sufficiently large, x_s^c is effectively removed from the regression model, see Hall, Li and Racine (2007) on a more detailed discussion on removing irrelevant covariates by oversmoothing these variables.

It is well known that the smoothing parameters play an essential role in the trade-off between reducing bias and variance, so that their choice in a nonparametric approach is very critical. For the aforementioned setting, Hall, Li and Racine (2007) suggested choosing the smoothing parameters $(h, \lambda) = (h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q)$ by minimizing the following cross-validation (CV) function:

$$CV(h, \lambda) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{g}_{-i}(X_i))^2 \, w_1(X_i), \tag{2.3}$$

where $\widehat{g}_{-i}(X_i) = \sum_{j \neq i}^{n} Y_j K(X_i, X_j) / \sum_{j \neq i}^{n} K(X_i, X_j)$ is the leave-one-out kernel estimator of $g(X_i) \equiv E(Y_i|X_i)$, and $0 \leq w_1(\cdot) \leq 1$ is a weight function (which has a compact support) that serves to avoid difficulties caused by dividing by zero, or by the slower convergence rate arising when X_i lies near the boundary of the support of X. Although it is necessary to introduce the weight function $w_1(\cdot)$ from the theoretical point of view, in practice the use of the weight function may not be necessary. In applications, since the data range is always finite, one usually does not need to use any weight function, or equivalently one can use $w_1(X_i) \equiv 1$ for all $i = 1, ..., n$.

Now suppose that X_s^d, the s-th component of X^d, is an irrelevant component, i.e., $E(Y_i|X_i = x) = E(Y_i|X_i/X_{is}^d = x/x_s^d)$ almost everywhere, where X_i/X_{is}^d denote the set of variables in X_i with X_{is}^d being removed. Let λ_s denote the smoothing parameter associated with irrelevant component X_s^d. Hall, Li and Racine (2007) showed that, when X_s^d is an irrelevant regressor, the cross-validated λ_s converges to 1 in probability. Recall that when $\lambda_s = 1$, the corresponding variable X_s^d is completely removed from the nonparametric kernel estimator $\widehat{g}(x)$. This means that all irrelevant discrete variables can be automatically removed (asymptotically) by the least squares cross-validation method. Similar results hold true for the continuous covariates. Indeed, Hall, Li and Racine (2007) showed that, when X_s^c is an irrelevant covariate, then the cross-validated smoothing parameter h_s diverges to $+\infty$. In such a case, the corresponding kernel function $w((X_{is}^c - x_s^c)/h_s) \to w(0)$ becomes a constant. Moreover, this constant is cancelled out from $\widehat{g}(x)$ because the same constant appears at both the numerator and the denominator of $\widehat{g}(x)$. Hence, asymptotically all irrelevant covariates, either continuous or discrete, is smoothed out from the regression model by the cross-validation method.

The nonparametric estimator $\widehat{g}(x)$ with the cross-validated smoothing parameters has the same asymptotic distribution of a kernel estimator of $g(x)$ that first removes the irrelevant covariates. Hall, Li and Racine (2007) defined the irrelevant variables as those regressors that are independent with both the dependent variable and the relevant regressors. However, the simulation results suggest that the cross-validation method can still remove irrelevant variables as long as those irrelevant variables are independent with the dependent variable conditional on the relevant variables. However, it is still of theoretical interest if one can also relax the independent assumption to conditional independent assumption, and this remains an interesting open question.

Note that the above result was extended by Li and Racine (2009) to the case of estimating a varying coefficient model and by Li, Ouyang and Racine (2009) and Su, Chen and Ullah (2009) to weakly dependent data case.

When all the covariates are discrete, the asymptotic analysis is quite different and cannot be obtained from the regression model with mixed discrete and continuous regressors as a special case (since the above result assumes that $q \geq 1$, where q is the number of continuous regressors). When all the regressors are discrete variables, irrelevant discrete covariates is smoothed out by the least squares cross-validation method with a positive probability, say δ. Indeed, Ouyang, Li and Racine (2009) concluded that $0.5 < \delta < 1$. More precisely, the simulation results reported in their paper suggest that $\delta \in [0.6, 0.65]$. In summary, when all the regressors are discrete, one can still remove the irrelevant regressors (by the cross-validation method) with a positive probability, but this probability is strictly less than one, even as the sample size goes to $+\infty$.

Finally, various programs for implementing the cross-validation method to estimate a regression model with mixed discrete and continuous covariates are available. For example, a R-package (np) is currently available at http://www.R-project.org for a free download and a Stata program will be available soon.

2.2 Consistent Model Specification Tests

It is well known that the selection of smoothing parameter is of crucial importance in nonparametric estimation. It is probably less well known (say, to applied econometricians) what important roles the smoothing parameters play in nonparametric model specification testing. In this subsection, we first consider a simple univariate regression model to illustrate how the selection of smoothing parameter affects the performance of a nonparametric test. Toward this end, we consider the following nonparametric regression model

$$Y_i = g(X_i) + u_i,$$

where X_i is a univariate continuous random variable and $g(\cdot)$ is a smooth function. We are interested in testing the null hypothesis $H_0 : E(Y_i|X_i) = \beta_0 + X_i\beta_1$ almost surely (a.s.). One can construct a test based on $I = E[u_i E(u_i|X_i)f(X_i)]$, where $u_i = Y_i - \beta_0 - X_i\beta_1$ and $f(\cdot)$ is the density function of X_i. This is because $I = E[(E(u_i|X_i))^2 f(X_i)] \geq 0$, and it equals to 0 if and only if the null hypothesis is true. Hence, I serves as a proper candidate for testing H_0. A feasible test statistic based on I is given by

$$I_n = \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i \widehat{E}_{-i}(u_i|X_i)\widehat{f}_{-i}(X_i) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} \widehat{u}_i\widehat{u}_j K_{h,ij},$$

where $K_{h,ij} = K_h(X_i - X_j)$ and $K_h(v) = h^{-1}K(v/h)$. It can be shown that I_n converges to 0 under H_0 (indeed, $I_n = O_p((nh^{1/2})^{-1})$ under H_0), and that I_n goes to a positive con-

stant if H_0 is false. A standardized test is given by $T_n = nh^{1/2} I_n / \widehat{\sigma}_0$, where $\widehat{\sigma}_0^2 = 2[n(n-1)h]^{-1} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \widehat{u}_i^2 \widehat{u}_j^2 K_{h,ij}^2$. One can show that T_n converges to a standard normal random variable under H_0, and it diverges to $+\infty$ at the rate of $nh^{1/2}$ if H_0 does not hold. In practice, some residual based bootstrap methods (say, the wild bootstrap method) are recommended for a better approximation to the finite sample null distribution of the test statistic T_n. The conditions on h are the usual ones: $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

Now the question is: how does the selection of h affect the performance of the T_n test? and how should we select h in practice? Given that residual based bootstrap methods can give quite satisfactory estimated sizes for T_n, a sensible starting point seems to examine the power property of the test. For a given significance level for a test, one would prefer a test with a large power. To examine how h affects the power of the test, we need to know the behavior of $g(x) \equiv E(Y_i | X_i = x)$ when H_0 fails to hold. In this case $g(x)$ is a nonlinear function of x. Let us consider a specific example. Suppose that $X \in [0, 2]$ and $g(x) = \sin(m\pi x)$, where m is a positive constant. Now consider the case that m is small, say $m = 1/4$. Then $g(x)$ changes from $\sin(0) = 0$ to $\sin(\pi/2) = 1$ as x varies from 0 to 2. The function is monotonically increasing (slowly) over the domain of x. For such a slowly changing function (as x varies), intuitively it is not hard to imagine that the optimal smoothing should be relatively large. In contrast, if $m = 2$, then $m\pi x$ changes from 0 to 4π (as x moves from 0 to 2) and the function $\sin(m\pi x)$ completes two full periods, moving up and down several times as x varies in the domain. This function changes more rapidly compared to the case of $m = 1/4$, the optimal smoothing for this fast changing function should be much smaller compared to a slow changing function (the case of $m = 1/4$). We generate X_i's uniformly from $[0, 2]$ and use the least squares cross-validation method to select the smoothing parameters. For a sample size of $n = 100$ and over $1,000$ simulations, the median value of $\widehat{h}$ (cross-validated h) is 0.172 for $m = 1/4$, and 0.068 for $m = 2$. If we use an ad-hoc rule such as $h = x_{sd} n^{-1/5} = 0.230$ for $n = 100$, where x_{sd} is the sample standard error of $\{X_i\}_{i=1}^n$. We say that the optimal smoothing parameter (in estimation) can be quite different depending on the different shapes of the unknown regression functions.

How is the nonparametric estimation accuracy related to a power of a nonparametric test? In general, more accurate estimation of the unknown function is expected to lead to a better power of a test if the test is based on the difference between the null hypothesized linear model and the true unknown function.[1] For this reason Hsiao, Li and Racine (2007) suggested using the least squares cross-validation method to select the smoothing parameters in a nonparametric smoothing test. Hsiao, Li and Racine (2007) considered the problem of testing a parametric

[1]This argument may not be always true as one can also choose a fixed value of h in testing problems, resulting in a non-smoothing test, see Chapter 13 of Li and Racine (2007) on more detailed discussions of non-smoothing tests.

regression functional form with mixed discrete and continuous covariates. We next describe their testing procedure.

For testing the null hypothesis that a parametric regression model is correctly specified, we state it as

$$H_0 : P[E(Y_i|X_i) = m(X_i, \beta)] = 1 \text{ for some } \beta \in \mathcal{B}, \tag{2.4}$$

where $m(\cdot, \cdot)$ is a known function with β being a $p \times 1$ vector of unknown parameters and $\mathcal{B}$ is a compact subset in $\Re^p$. The alternative hypothesis is the negation of H_0, i.e.,

$$H_1 : P[E(Y_i|X_i) = m(X_i, \beta)] < 1 \text{ for all } \beta \in \mathcal{B}. \tag{2.5}$$

Hsiao, Li and Racine (2007) considered a test statistic that was independently proposed by Fan and Li (1996) and Zheng (1996). The test statistic is based on I=$E[u_i E(u_i|X_i)f(X_i)]$ as we discussed earlier. The sample analogue of I is given by

$$\begin{aligned} I_n &= n^{-1}\sum_{i=1}^{n} \widehat{u}_i \widehat{E}_{-i}(u_i|X_i)\widehat{f}_{-i}(X_i) = n^{-1}\sum_{i=1}^{n} \widehat{u}_i \left\{ n^{-1} \sum_{j=1, j\neq i}^{n} \widehat{u}_j W_{h,ij} L_{\lambda,ij} \right\} \\ &= n^{-2}\sum_{i}\sum_{j\neq i} \widehat{u}_i \widehat{u}_j K_{\gamma,ij}, \end{aligned} \tag{2.6}$$

where $K_{\gamma,ij} = W_{h,ij} L_{\lambda,ij}$ ($\gamma = (h, \lambda)$), $\widehat{u}_i = Y_i - m(X_i, \widehat{\beta})$ is the residual obtained from estimating the parametric null model, $\widehat{\beta}$ is a $\sqrt{n}$-consistent estimator of β (under H_0), and $\widehat{E}_{-i}(u_i|X_i)\widehat{f}_{-i}(X_i)$ is a leave-one-out kernel estimator of $E(Y_i|X_i)f(X_i)$. In the case where we have only continuous regressors X_i^c and use a non-stochastic value of h_s ($h_s \to 0$ and $nh_1 \dots h_q \to \infty$), the asymptotic null (normal) distribution of the I_n test was derived independently by Fan and Li (1996) and Zheng (1996).

For the I_n test with the mixed discrete and continuous covariates, Hsiao, Li and Racine (2007) advocated the use of cross-validation methods for selecting the smoothing parameter vectors h and λ. We use $\widehat{I}_n$ to denote the test statistic with CV selected smoothing parameters, i.e, $\widehat{I}_n$ is defined the same way as I_n given in (2.6) but with $(h_1, \dots, h_q, \lambda_1, \dots, \lambda_r)$ replaced by the CV smoothing parameters $(\widehat{h}_1, \dots, \widehat{h}_q, \widehat{\lambda}_1, \dots, \widehat{\lambda}_r)$. The asymptotic distribution of our CV-based test was derived by Hsiao, Li and Racine (2007):

$$\widehat{T}_n \equiv n(\widehat{h}_1 \dots \widehat{h}^q)^{1/2}\widehat{I}_n/\sqrt{\widehat{\Omega}} \xrightarrow{d} N(0, 1)$$

under H_0, where "$\xrightarrow{d}$" denotes the convergence in distribution and $\widehat{\Omega} = [2(\widehat{h}_1 \dots \widehat{h}_q)/n^2] \sum_{i=1}^{n} \sum_{j\neq i}^{n} \widehat{u}_i^2 \widehat{u}_j^2 W_{\widehat{h},ij}^2 L_{\widehat{\lambda},ij}^2$.

Hsiao, Li and Racine (2007) also showed that the $\widehat{T}_n$ test diverges to $+\infty$ if H_0 is false; thus it is a consistent test. Hsiao, Li and Racine (2007) recommended the use of a residual-based wild bootstrap method to better approximate the null distribution of $\widehat{T}_n$. Specifically, one generates

the wild bootstrap error u_i^* via a two point distribution $u_i^* = [(1-\sqrt{5})/2]\widehat{u}_i$ with probability $(1+\sqrt{5})/[2\sqrt{5}]$, and $u_i^* = [(1+\sqrt{5})/2]\widehat{u}_i$ with probability $(\sqrt{5}-1)/[2\sqrt{5}]$. Using $\{u_i^*\}_{i=1}^n$, one generates $Y_i^* = m(X_i, \widehat{\beta}) + u_i^*$ for $i = 1, \ldots, n$. $\{X_i, Y_i^*\}_{i=1}^n$ is called the 'bootstrap sample', and one uses this bootstrap sample to obtain a nonlinear least squares estimator of β (a least squares estimator if $m(X_i, \beta) = X_i^T\beta$). Let $\widehat{\beta}^*$ denote the resulting estimator. The bootstrap residual is given by $\widehat{u}_i^* = Y_i^* - m(X_i, \widehat{\beta}^*)$. The bootstrap test statistic $\widehat{T}_n^*$ is obtained the same way as $\widehat{T}_n$ with $\widehat{u}_i$ being replaced by $\widehat{u}_i^*$. Note that we use the same CV selected smoothing parameters $\widehat{h}$ and $\widehat{\lambda}$ when computing the bootstrap statistics. That is, there is no need to rerun CV with the bootstrap sample. Therefore, our bootstrap test is computationally quite simple. In practice, one repeats the above steps a large number of times, say $B = 1000$ times, then, the original test statistic $\widehat{T}_n$ plus the B bootstrap test statistics give us the empirical distribution of the bootstrap statistics, which is then used to approximate the finite-sample null distribution of $\widehat{T}_n$.

By adopting the concept of 'convergence in distribution in probability' (e.g., Li, Hsiao and Zinn (2003)) to study the asymptotic distribution of the bootstrap statistic $\widehat{T}_n^*$, Hsiao, Li and Racine (2007) showed that the wild bootstrap method works by proving the following result

$$\sup_{z\in\Re} |P(\widehat{T}_n^* \leq z | \{X_i, Y_i\}_{i=1}^n) - \Phi(z)| = o_p(1), \tag{2.7}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. The simulation results reported in Hsiao, Li and Racine (2007) show that the proposed bootstrap procedure indeed works well in finite sample applications. See Hsiao, Li and Racine (2007) for details on this regard.

2.3 Testing Significance (Relevance) of Discrete Variables

When all the regressors are discrete variables, Ouyang, Li and Racine (2009) showed that while the irrelevant variables can be smoothed out with about 65% probability, there is a 35% probability that the cross-validated λ takes values strictly less than 1 even as $n \to \infty$. Therefore, sometimes the cross-validation method may not be able to determine whether a given variable is irrelevant or not. In such cases, one can use the test statistic proposed by Racine, Hart and Li (2006) to test whether a given discrete variable is relevant or not. The null hypothesis is

$$H_0: \qquad m(x,z) = E(Y|X=x, Z=z) = E(Y|X=x) \text{ almost everywhere (a.e.)}, \tag{2.8}$$

where Z is a discrete variable and X can contain both discrete and continuous components. Under the null hypothesis, the discrete variable Z is an irrelevant regressor.

Assume that Z takes c different values, without loss of generality, say that $Z \in \{0, 1, ..., c-1\}$. The null hypothesis H_0 is equivalent to: $m(X, Z=l) = m(X, Z=0)$ for $l = 1, \ldots, c-1$ (for all

X). Racine, Hart and Li (2006) suggested constructing a test statistic based on

$$I = \sum_{l=1}^{c-1} E\left\{[m(X, Z = l) - m(X, Z = 0)]^2\right\}. \tag{2.9}$$

Obviously, $I \geq 0$ and $I = 0$ if and only if H_0 is true. Therefore, I serves as a proper measure for testing H_0. A feasible test statistic is given by

$$\widehat{I}_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{c-1}[\widehat{m}(X_i, Z_i = l) - \widehat{m}(X_i, Z_i = 0)]^2, \tag{2.10}$$

where $\widehat{m}(X_i, Z_i)$ is the kernel estimator of $m(X_i, Z_i)$.

Racine, Hart and Li (2006) recommended using the least squares cross-validation method to select the smoothing parameters. Let $\widehat{\lambda}_z$ denote the smoothing parameter selected by the cross-validation method. Since under H_0, $\widehat{\lambda}_z$ has a non-degenerate (complicated) limiting distribution, the null distribution of $\widehat{I}_n$ is unknown even as $n \to \infty$. Therefore, Racine, Hart and Li (2006) recommended using some bootstrap procedures to approximate the null distribution of the $\widehat{I}_n$ test, one of which is described below.

A Bootstrap Procedure

1. Randomly select Z_i^* from $\{Z_j\}_{j=1}^n$ with replacement, and call $\{Y_i, X_i, Z_i^*\}_{i=1}^n$ the bootstrap sample.
2. Use the bootstrap sample to compute the bootstrap statistic $\widehat{I}_n^*$, where $\widehat{I}_n^*$ is the same as $\widehat{I}_n$ except that Z_i is replaced by Z_i^* (using the same cross-validated smoothing parameters of $\widehat{h}$, $\widehat{\lambda}$ and $\widehat{\lambda}_z$ obtained earlier).
3. Repeat steps 1 and 2 a large number of times, say B times. Let $\{\widehat{I}_{n,j}^*\}_{j=1}^B$ be the ordered (in an ascending order) statistic of the B bootstrap statistics, and let $\widehat{I}_{n,(\alpha)}^*$ denote the $(1-\alpha)$th percentile of $\{\widehat{I}_{n,j}^*\}_{j=1}^B$. We reject H_0 if $\widehat{I}_n > \widehat{I}_{n,(\alpha)}^*$ at the level α.

The simulation results reported in Racine, Hart and Li (2006) show that the above bootstrap procedure works well in finite sample applications. See Racine, Hart and Li (2006) for details on empirical studies.

3 Nonparametric Regression Models With Nonstationary Data

Phillips and Park (1998) were the first to study the asymptotic theory on nonparametric estimation of econometric models with nonstationary data. Recently, nonparametric estimation of regression functions has attracted many attentions among statisticians and econometricians.

Juhl (2005) and Wang and Phillips (2006, 2008) considered nonparametric regression models with nonstationary regressors, while Cai, Li and Park (2009) and Xiao (2009) considered semiparametric varying coefficient models with some of the regressors being nonstationary. Gao, King, Lu and Tjøstheim (2008) and Sun, Cai and Li (2008) considered nonparametric testing issues with nonstationary data. Finally, Karlsen, Myklebust and Tjøstheim (2007) considered nonparametric estimation of a regression model for a more general type of nonstationary processes, a subclass of the class of null recurrent Markov chains. We summarize some of these works below.

3.1 Nonparametric Density And Regression Function Estimation

Phillips and Park (1998) considered a nonparametric autoregressive regression model with the true data generated by an unit root process:

$$Y_t = m(Y_{t-1}) + u_t \equiv Y_{t-1} + u_t,$$

where u_t, for expositional simplicity, is assumed to be i.i.d. $(0, \sigma_u^2)$. Phillips and Park (1998) suggested using a local constant method to estimate $m(\cdot)$ as

$$\widehat{m}(x) = \frac{\sum_{t=1}^n Y_t K_h(Y_{t-1} - x)}{\sum_{t=1}^n K_h(Y_{t-1} - x)} \equiv \frac{(nh)^{-1} \sum_{t=1}^n Y_t K_h(Y_{t-1} - x)}{\widehat{f}_n(x)}, \tag{3.1}$$

where $K_h(v) = h^{-1}K(v/h)$, h is the bandwidth, $K(\cdot)$ is the kernel function, and $\widehat{f}_n(x) = (nh)^{-1} \sum_{t=1}^n K_h(Y_{t-1} - x)$, which would be regarded as an estimator of the density function if Y_t were stationary. Phillips and Park (1998) derived the asymptotic distributions for both $\widehat{m}(x)$ and $\widehat{f}_n(x)$.

It follows from Donsker's theorem that under some regularity conditions, for $0 \leq r \leq 1$, $Y_{[nr]}/\sqrt{n} \implies W_u(r)$, where $[\cdot]$ denotes the integer part of $\cdot$, $\Longrightarrow$ denotes weak convergence, $W_u(\cdot)$ is a Brownian motion on $[0,\ 1]$, $\sigma_u^{-1} W_u(r)$ is a standard Brownian motion on $[0,\ 1]$, and $\sigma_u^2 = E(u_t^2)$. Define the local time $L_W(t,x)$ for a Brownian motion $W(\cdot)$ as

$$L_W(t,x) = \lim_{\epsilon \to 0} \frac{1}{2\epsilon} \int_0^t \mathbf{1}(|W(s) - x| \leq \epsilon) ds. \tag{3.2}$$

Under some regularity conditions including $h \to 0$ and $nh \to \infty$ as $n \to \infty$, Phillips and Park (1998) established the following result:

$$n^{1/4} h^{1/2} (\widehat{m}(x) - m(x)) \xrightarrow{d} MN\left(0,\ \sigma_u^2 \nu_0(K) / L_{W_u}(1,0)\right), \tag{3.3}$$

where $MN(\mu, \Sigma)$ denotes a mixed normal distribution with mean μ and conditional variance Σ, and $\nu_0(K) = \int K^2(v) dv$. Note that there is no bias term in (3.3) because $m(x) = x$ is a linear function so that its derivatives with orders greater or equal to two all vanish.

Wang and Phillips (2006) considered the following nonlinear cointegration model:

$$Y_t = g(X_t) + u_t, \qquad t = 1, 2, ..., n,$$

where $X_0 = 0$ and $X_t = X_{t-1} + \epsilon_t$, both u_t and ϵ_t are mean zero stationary processes. Wang and Phillips (2006) considered the local constant estimator for $g(x)$ given by

$$\widehat{g}(x) = \frac{\sum_{t=1}^{n} Y_t K_h(X_t - x)}{\sum_{t=1}^{n} K_h(X_t - x)}.$$

Under some regularity conditions including $nh \to \infty$ and $nh^3 \to 0$ (undersmoothing) as $n \to \infty$, Wang and Phillips (2006) showed that

$$\left(n^{-1/2} \sum_{t=1}^{n} K_h(X_t - x) \right)^{1/2} n^{1/4} h^{1/2} \left(\widehat{g}(x) - g(x) \right) \xrightarrow{d} N(0, \sigma_1^2), \tag{3.4}$$

where $\sigma_1^2 = \sigma_u^2 \nu_0(K)$. When $X_t = Y_{t-1}$, (3.4) gives the asymptotic distribution of $\widehat{m}(x)$ defined in (3.1). This is because the asymptotic variances in (3.3) and (3.4) are the same since it can be shown that $n^{-1} \sum_{t=1}^{n} K_h(X_t - x) \xrightarrow{p} L_W(1, 0)/\sigma_\epsilon$, where $W(\cdot)$ is a standard Brownian motion and $\sigma_\epsilon^2 = \lim_{n\to\infty} Var(n^{-1/2} \sum_{t=1}^{n} \epsilon_t)$ ($\sigma_\epsilon^2 = Var(\epsilon_t)$ if ϵ_t is serially uncorrelated). Finally, Wang and Phillips (2008) extended the result of Wang and Phillips (2006) to allow for endogenous regressors.

3.2 Semiparametric Estimation Of A Varying Coefficient Model With Nonstationary Covariates

Cai, Li and Park (2009) considered the following varying coefficient model

$$Y_t = X_t^T \beta(Z_t) + u_t = X_{t1}^T \beta_1(Z_t) + X_{t2}^T \beta_2(Z_t) + u_t, \qquad t = 1, ..., n, \tag{3.5}$$

where A^T denotes the transpose of a matrix or vector A, X_{t1}, Z_t, and u_t are stationary, X_{t2} is an I(1) process, $\beta(Z_t) = (\beta_1(Z_t)^T, \beta_2(Z_t)^T)^T$, and $X_t = (X_{t1}^T, X_{t2}^T)^T$. Here X_{ti} is a $d_i \times 1$ vector, $i = 1, 2$, $d_1 + d_2 = d$, and the first component of X_{t1} is identically one. Also, Y_t, Z_t and u_t are scalars, and $E(u_t) = 0$, $\sigma_u^2 = \lim_{n\to\infty} Var(n^{-1/2} \sum_{t=1}^{n} u_t)$ is finite, and u_t is assumed to be independent with (X_t, Z_t).[2] When there is no term $X_{t1}^T \beta_1(Z_t)$, (3.5) reduces to the model investigated by Xiao (2009). Note that Y_t can be stationary or nonstationary. If Y_t is nonstationary, model (3.5) implies that Y_t and X_{t2} are co-integrated with a varying co-integration vector $\beta_2(Z_t)$. The reason why Cai, Li and Park (2009) considered a following varying coefficient model in (3.5) is that it might approximate a general nonparametric model well (see (4.8) for details).

[2]This independence assumption can be relaxed to $E(u_t|X_t, Z_t) = 0$, which leads to some modification to the asymptotic theory.

It is easy to see that the local linear estimator for $\beta(z)$ and its derivative function $\beta^{(1)}(z) = d\beta(z)/dz$ is given by

$$\begin{pmatrix} \widehat{\beta}(z) \\ \widehat{\beta}^{(1)}(z) \end{pmatrix} = \left[\sum_{t=1}^{n} \begin{pmatrix} X_t \\ (Z_t - z)\, X_t \end{pmatrix}^{\otimes 2} K_h(Z_t - z) \right]^{-1} \sum_{t=1}^{n} \begin{pmatrix} X_t \\ (Z_t - z) X_t \end{pmatrix} Y_t\, K_h(Z_t - z), \quad (3.6)$$

where $A^{\otimes 2} = AA^T$ and $A^{\otimes 1} = A$.

We assume that X_{t2} can be written as $X_{t2} - X_{t-1,2} = \eta_t$, where η_t is a zero mean stationary process. Then under some standard regularity conditions, $X_{t2}/\sqrt{n} \implies W_{\eta 2}(r)$, where $W_{\eta 2}(\cdot)$ is a d_2-dimensional Brownian motion on [0, 1]. By the continuous mapping theorem we know that, for $l = 1, 2$

$$\frac{1}{n} \sum_{t=1}^{n} \left(X_{t2}/\sqrt{n} \right)^{\otimes l} \xrightarrow{d} \int_0^1 [W_{\eta 2}(r)]^{\otimes l} dr \equiv W_{\eta 2}^{(l)}. \quad (3.7)$$

Let $f_z(z)$ be the marginal density of Z_t. Define $M_k(z) = E\left[X_{t1}^{\otimes k} \,|\, Z_t = z \right]$ for $1 \le k \le 2$. Further, let

$$S(z) = \begin{pmatrix} M_2(z) & M_1(z)\, W_{\eta 2}^{(1)\,T} \\ W_{\eta 2}^{(1)}\, M_1(z)^T & W_{\eta 2}^{(2)} \end{pmatrix},$$

and $D_n = \mathrm{diag}\{I_{d_1}, \sqrt{n}\, I_{d_2}\}$. Then, Cai, Li and Park (2009) showed that under some regularity conditions,

$$\sqrt{nh}\, D_n \left[\widehat{\beta}(z) - \beta(z) - \frac{1}{2}\, h^2\, \mu_2(K) \beta^{(2)}(z) \right] \xrightarrow{d} MN(0, \Sigma_\beta(z)), \quad (3.8)$$

where $MN(0, \Sigma_\beta(z))$ is a mixed normal variable with mean zero and conditional covariance $\Sigma_\beta(z) = \sigma_u^2 \nu_0(K) S(z)^{-1} / f_z(z)$ and $\mu_2(K) = \int v^2 K(v) dv$.

Equation (3.8) implies that $\widehat{\beta}_1(z) - \beta_1(z) = O_p(h^2 + (nh)^{-1/2})$ and $\widehat{\beta}_2(z) - \beta_2(z) = O_p(h^2 + (n^2 h)^{-1/2})$. Thus, the convergence rate for $\widehat{\beta}_2(z) - \beta_2(z)$ is faster than that of $\widehat{\beta}_1(z) - \beta_1(z)$. The bias term is $O(h^2)$ for both $\widehat{\beta}_1(z)$ and $\widehat{\beta}_2(z)$, and the variance of $\widehat{\beta}_1(z)$ is $O((nh)^{-1})$, while the variance of $\widehat{\beta}_2(z)$ is $O((n^2 h)^{-1/2})$. This is similar to the linear regression model case because $\sum_{t=1}^{n} X_{2t} X_{t2}^T = O_p(n^2)$ and $\sum_{t=1}^{n} X_{t1} X_{t1}^T = O_p(n)$. The estimated coefficient for the $I(1)$ regressor is n-consistent, while the estimated coefficient for the $I(0)$ regressor has the standard $\sqrt{n}$ rate of convergence.

Cai, Li and Park (2009) also considered the case that X_t is $I(0)$ but Z_t is $I(1)$. For such a case, Z_t can be expressed as $Z_t = Z_{t-1} + v_t = Z_0 + \sum_{s=1}^{t} v_s$, where $\{v_s\}$ is a stationary process with mean zero and $\sigma_v^2 = \lim_{n\to\infty} \mathrm{Var}(n^{-1/2} \sum_{t=1}^{n} v_t) > 0$. Then, it follows from Donsker's theorem that under some regularity conditions, for $0 \le r \le 1$, $Z_{[nr]}/\sqrt{n} \implies W_v(r)$, where $W_v(\cdot)$ is a Brownian motion on [0, 1] and $\sigma_v^{-1} W_v(r)$ is a standard Brownian motion on [0, 1]. Cai, Li and Park (2009) established the following asymptotic result:

$$\sqrt{n^{1/2} h} \left[\widehat{\beta}(z) - \beta(z) - h^2 B(z) \right] \xrightarrow{d} MN(0, \Sigma_1), \quad (3.9)$$

where $B(z) = \mu_2(K)\beta^{(2)}(z)/2$, $MN(0, \Sigma_1)$ is a mixed normal distribution with mean zero and conditional covariance $\Sigma_1 = \sigma_v \sigma_u^2 \, \nu_0(K) \, [E(X_t X_t^T) L_W(1,0)]^{-1}$. Equation (3.9) implies that $\widehat{\beta}(z) - \beta(z) = O_p(h^2 + (n^{1/4} h^{1/2})^{-1})$ so that the optimal smoothing h is proportional to $n^{-1/10}$. Thus, h should converge to 0 at a fairly slow rate at $n^{-1/10}$. This is because when Z_t is $I(1)$, it returns to the fixed interval $[z-h, z+h]$ less often compared to the case when Z_t is $I(0)$. Therefore, one needs to let h go to 0 slowly so as to balance the squared bias and the variance.

When $d = 1$ and $X_t \equiv 1$, the varying coefficient model reduces to a simple regression model $Y_t = \beta(Z_t) + u_t$ (Z_t is $I(1)$). The asymptotic variance in (3.9) simplifies to $\sigma_v \sigma_u^2 \nu_0(K) L_W(1,0)^{-1}$. It can be shown that $\widehat{f}(z) \equiv n^{-1/2} \sum_{t=1}^n K_h(Z_t - z)$ consistently estimates $L_W(1,0)/\sigma_v$; see Phillips and Park (1998). Hence, in this case (3.9) can be equivalently written as

$$[\widehat{\sigma}_u^2 \nu_0(K)]^{-1/2} [\widehat{f}(z)]^{1/2} \sqrt{n^{1/2} \, h} \, \left[\widehat{\beta}(z) - \beta(z) - h^2 B(z) \right] \xrightarrow{d} N(0,1), \tag{3.10}$$

where $\widehat{\sigma}_u^2 = n^{-1} \sum_{i=1}^n [Y_t - \widehat{\beta}(Z_t)]^2$ is a consistent estimator for σ_u^2. As expected, (3.10) is the same as that in Wang and Phillips (2006) for a nonparametric regression model with an $I(1)$ regressor.

Bachmeier, Leelahanon and Li (2006) considered the following semiparametric dynamic varying coefficient model:

$$Y_t = \beta_1(Z_t) + Y_{t-1} \beta_2(Z_t) + u_t, \tag{3.11}$$

where Y_t is the rate of inflation, and Z_t is an I(1) variable 'velocity of money supply'. Bachmeier, Leelahanon and Li (2006) applied the above model to forecast U.S. inflation rate and showed that the semiparametric varying coefficient dynamic model (with a nonstationary covariate) has smaller forecast mean squared error compared with the conventional linear model, or some nonparametric model using only stationary covariates. For more examples in finance, the reader is referred to the paper by Cai and Hong (2009).

Park and Hahn (1999) considered the varying coefficient model in (3.5) with Z_t being replaced by the time trend variable t, and established the asymptotic distribution of a series-based estimator for $\beta(t)$. Park and Hahn (1999) also proposed a test statistics for testing a parametric function form for $\beta(\cdot)$ and for testing co-integration in a time varying coefficient model framework.

Cai and Wang (2009) considered a similar time varying coefficient model as the one considered in Park and Hahn (1999) with nonstationary or nearly nonstationary (local to unit root) and endogenous regressors. Cai and Wang (2009) used a local linear estimation method and derived the asymptotic distribution of their proposed estimators. Finally, Cai and Wang (2009) applied the above model to test the stability of the predictability of asset returns in finance. That is,

$$r_t = \beta_{0t} + \beta_{1t} \, x_{t-1} + u_t,$$

where r_t is the asset return and x_{t-1} is the first lag of financial instrument, say the logarithm of the earnings-price ratio or the dividend-price ratio or other financial variables. But u_t and x_{t-1} is usually correlated and x_t is nonstationary like I(1) or near I(1) and highly persistent. For details about the theory and applications, we refer the reader to the paper by Cai and Wang (2009).

3.3 Data-Driven Method Of Selecting Smoothing Parameter

Sun and Li (2009) considered the problem of selecting the smoothing parameter h of model (3.5) by the least squares cross-validation method. They proposed to choosing h by minimizing the following least squares cross-validation objective function:

$$CV(h) = n^{-1} \sum_{t=1}^{n} \left[Y_t - X_t^T \widehat{\beta}_{-t}(Z_t) \right]^2 M(Z_t), \tag{3.12}$$

where $\widehat{\beta}_{-t}(Z_t)$ is a leave-one-out kernel estimator of $\beta(Z_t)$.

Sun and Li (2009) first considered the case that X_t is I(1) (there is no I(0) components in X_t), Z_t and u_t are stationary processes. They found an interesting result that the local constant (LC) and the local linear (LL) estimation methods lead to very different asymptotic behaviors for $\widehat{h}$ by the CV method selected smoothing parameter. Specifically, they showed that for the local constant estimation method (assuming X_t is a scalar to simplify the notation)

$$\sqrt{n}\,\widehat{h}_{lc-cv} - \sqrt{\frac{c_{1n}\sigma_u^2\nu_0 \int M(z)\,dz}{\nu_2 c_{2n} \int \left(\beta^{(1)}(z)\right)^2 M(z)\,dz}} \xrightarrow{p} 0, \tag{3.13}$$

where $c_{1n} = n^{-2}\sum_{t=1}^{n} X_t^2$, $c_{2n} = n^{-3}\sum_t X_t^4$, and $\nu_j = \int v^j K^2(v)\,du$. For the local linear estimation method the result is

$$n^{2/5}\widehat{h}_{ll-cv} - \left(\frac{4\sigma_u^2\nu_0 \int M(z)\,dz}{c_{1n}\mu_2(K) E\left(\left(\beta_t^{(2)}\right)^2 M_t \right)} \right)^{1/5} \xrightarrow{p} 0. \tag{3.14}$$

One interesting implication of (3.13) and (3.14) is that the CV selected h is stochastic even asymptotically. Also, comparing (3.13) with (3.14) we see that the CV selected h has different convergence rates. Both these results are in sharp contrast to the stationary data or independent data case where we know that the CV selected smoothing parameter is asymptotically non-stochastic and that the CV functions have the same probability order whether one uses the LC or the LL method. The reason for the different rates of convergence of $\widehat{h}$ is that $CV_{LC}(h) = O_p(h + (nh)^{-1})$, while $CV_{LL}(h) = O_p(nh^4 + (nh)^{-1})$. This also implies that $CV_{LC}(\widehat{h}) = O_p(n^{-1/2})$ and $CV_{LL}(\widehat{h}) = O_p(n^{-3/5})$. Hence, the LL method leads to more efficient estimation than the LC method.

Sun and Li (2009) further provided asymptotic analysis for CV selected h for model (3.5) with X_t containing both I(0) and I(1) components.

3.4 Testing A Parametric Coefficient Functional Form

Sun, Cai and Li (2008) considered the problem of testing the null hypothesis (H_0) that $P(\beta(Z) = \beta_0) = 1$ for some $d \times 1$ vector of constant coefficient β_0 in the following semiparametric model:

$$Y_t = X^T\beta(Z_t) + u_t = X_{1t}^T\beta_1(Z_t) + X_{2t}^T\beta_2(Z_t) + u_t,$$

where X_{1t}, Z_t and u_t are $I(0)$ variables, and X_{2t} is an $I(1)$ process. They proposed a test statistic based on the sample analogue of $\int ||D(\widehat{\beta}(z) - \widehat{\beta}_0(z))||^2 dz$, where $\widehat{\beta}(z)$ is the semi-parametric estimator of $\beta(z)$, $\widehat{\beta}_0$ is the least squares estimator of β_0 and D is a positive definite weight matrix. The test statistic proposed by Sun, Cai and Li (2008) can be simplified to

$$\widehat{I}_n = \frac{1}{n^3}\sum_{t=1}^{n}\sum_{s\neq t}^{n} X_t^T X_s \widehat{u}_t\widehat{u}_s K_{h,ts}, \tag{3.15}$$

where $\widehat{u}_t$ is the residual obtained from the parametric null model.

Sun, Cai and Li (2008) showed that under some regularity conditions and under H_0,

$$\widehat{J}_n = n\sqrt{h}\,\widehat{I}_n/\sqrt{\widehat{\sigma}_b^2} \xrightarrow{d} N(0,1),$$

where $\widehat{\sigma}_b^2 = n^{-4}h\sum_{t=1}^{n}\sum_{s\neq t}^{n}\tilde{u}_t^2\tilde{u}_s^2\left[X_t^T X_s\right]^2 K_{h,ts}^2$, $\tilde{u}_t = Y_t - X_t^T\widehat{\beta}_{-t}(Z_t)$ is the nonparametric residual and $\widehat{\beta}_{-t}(Z_t)$ is the leave-one-out estimator of $\beta(Z_t)$.

The power of the test statistic J_n depends on whether $\beta_2(z) = \beta_{20}$ or not, where β_{20} is a vector of constant parameters. If $\beta_2(z) \neq \beta_{20}$ for some z in a set with positive measure, Sun, Cai and Li (2008) showed that the $\widehat{J}_n$ test statistic diverges to $+\infty$ at the rate of n^2h. However, when $\beta_2(z) = \beta_{20}$ for all z, and $\beta_1(z) \neq \beta_{10}$ on a set with positive measure, $\widehat{J}_n$ diverges to $+\infty$ at the rate of $n\sqrt{h}$. Intuition behind this result is that, since $X_{2t}X_{2t}^T$ is larger than $X_{1t}X_{1t}^T$ by an order of n, hence, the test statistic diverges to $+\infty$ at a faster rate when $\beta_2(z)$, the coefficient of X_{2t}, is not a constant vector. We summarize the above results on power of the J_n test statistic as follow.

Sun, Cai and Li (2008) showed that under some regularity conditions and H_1, the following two results hold.

(i) If $P[\beta_2(Z_t) = \beta_{20}] < 1$ for any $\beta_{20} \in \mathcal{B}_2$, where $\mathcal{B}_2$ is a compact subset of $\mathcal{R}^{d_2}$, then $P[J_n > B_n] \to 1$ as $n \to \infty$ for any non-stochastic sequence $B_n = o(n^2\sqrt{h})$.

(ii) If $P[\beta_2(Z_t) = \beta_{20}] = 1$ for some $\beta_{20} \in \mathcal{B}_2$, and $P[\beta_1(Z_t) = \beta_{10}] < 1$ for any $\beta_{10} \in \mathcal{B}_1$, where $\mathcal{B}_1$ is a compact subset of $\mathcal{R}^{d_1}$, then $P[J_n > B_n] \to 1$ as $n \to \infty$ for any non-stochastic sequence $B_n = o(n\sqrt{h})$.

The above results imply that under H_1, the test statistic J_n diverges to $+\infty$ at different rates depending on whether $\beta_2(z) = \beta_{20}$ (a constant vector) or not. Nevertheless, the test statistic J_n is consistent in both cases, and a larger sample size might be required for the power of the test statistic to approach one if $\beta_2(z) = \beta_{20}$, and only the coefficients associated with the $I(0)$ variables are non-constant ($\beta_1(z) \neq \beta_{10}$).

Also, Sun, Cai and Li (2008) showed that when $\beta_1(z) = \beta_{10}$ (a constant vector) for all z, and $\beta_2(z) \neq \beta_{20}$, then the least squares estimator $\widehat{\beta}_{10}$ diverges to $+\infty$ at the rate of $\sqrt{n}$. Therefore, a misspecified linear model not only leads to inconsistent estimation result but also over-estimates the true parameter β_{10} by a different order of magnitude (the true $\beta_{10} = O(1)$ is finite, while $\widehat{\beta}_{10}$ diverges to ∞ at the rate of $\sqrt{n}$). Thus, one drastically over-estimates β_{10} in such a case if one estimates a misspecified linear model in which one assumes that the model is linear in both X_{1t} and X_{2t}, while in fact the true model is only linear in stationary covariate X_{1t}, but the coefficient of the nonstationary variable X_{2t} is a smoothing function of the stationary covariate Z_t. This result suggests that it is very important to test if the model specification is correct when there are integrated regressors in the model.

3.5 Testing Co-Integration in Semiparametric Varying Coefficient Models

In this subsection, we discuss the problem of testing whether u_t is an I(1) or an I(0) process through a varying coefficient model:

$$Y_t = X_t^T \beta(Z_t) + u_t,$$

where X_t is a $d \times 1$ vector of $I(1)$ variables, Z_t is an $I(0)$ scalar process, and u_t follows an AR(1) process as

$$u_t = \rho u_{t-1} + \epsilon_t,$$

where ϵ_t is a mean zero stationary process.

Xiao (2009) set the null hypothesis as H_0^a: u_t is an I(0) process (i.e., $\rho = 0$) and the alternative is H_1^a: u_t is an I(1) process ($\rho = 1$). It is easy to see that under H_0^a, $\text{Var}(u_t) = \sigma_u^2$, a positive constant, while under H_1^a, $\text{Var}(u_t) = a_0 + a_1 t$, where a_0 and a_1 are positive constants. Hence, Xiao (2009) suggested testing H_0^a by testing $a_1 = 0$. The test statistic is based on the following regression:

$$\widehat{u}_t^2 = a_0 + a_1 t + \text{ error}, \tag{3.16}$$

where $\widehat{u}_t = Y_t - X_t^T \widehat{\beta}(Z_t)$. Xiao (2009) showed that under H_0^a, $\widehat{t}_{a_1} = \widehat{a}_1 / se(\widehat{a}_1) \xrightarrow{d} N(0,1)$, where $\widehat{a}_1$ is the OLS estimator of a_1 based on (3.16) and $se(\widehat{a}_1)$ is the estimated standard error of $\widehat{a}_1$.

However, Sun and Li (2009) considered the case that under the null hypothesis, u_t is an I(1) process. Therefore, the null hypothesis considered by Sun and Li (2009) is H_0^b: u_t is an I(1) process, and the alternative is H_1^b: u_t is an I(0) process. Thus, the null hypothesis is H_0^b: $\rho = 1$ and the alternative hypothesis is H_1^b: $|\rho| < 1$. We consider only the case that $\beta(z)$ is not a constant function. Based on the well established cointegration testing for linear models, one can test H_0 based on

$$\widehat{\rho} = \frac{\sum_t \widehat{u}_t \widehat{u}_{t-1}}{\sum_t \widehat{u}_{t-1}^2},$$

where $\widehat{u}_t$ is an estimator for $u_t = Y_t - X_t^T \beta(Z_t)$ and the test statistic is $n(\widehat{\rho} - 1)$. Sun and Li (2009) showed that the leading term of the test statistic depends on $\widehat{\beta}(Z_t)$ in a complicated way and the asymptotic distribution is not nuisance parameter free. Therefore, one needs to design some simulation (or bootstrap) methods to approximate the null distribution of $n(\widehat{\rho} - 1)$. It is still an open question as how to approximate the null distribution of the test statistic considered by Sun and Li (2009).

3.6 Varying Coefficient Models With Time Trend Variables

Gu and Hernandez-Verme (2009) and Liang and Li (2009) considered a varying coefficient model with regressors containing a time trend:

$$Y_t = X_t^T \beta(Z_t) + u_t, \tag{3.17}$$

where $X_t^T = (X_{1t}^T, t)$ and X_{1t} is an I(0) variable. Gu and Hernandez-Verme (2009) considered the local linear estimation method and applied the method to evaluate the presence of credit rationing in the U.S. credit markets, while Liang and Li (2009) considered both the local constant and local polynomial estimation methods.

3.7 Varying Coefficient Models with I(1) Error

Sun, Hsiao and Li (2008) consider the problem of estimating a varying coefficient model

$$Y_t = X_t^T \beta(Z_t) + u_t, \tag{3.18}$$

when both X_t and the error term u_t are integrated I(1) processes. They show that in this case it is still possible to obtain consistent estimate of $\beta(\cdot)$, but the rate of convergence will be reduced to $O_p(h^2 + (nh)^{-1/2})$ rather than $O_p(h^2 + (n^2h)^{-1/2})$ as compared to the case when u_t is a stationary process.

4 Nonparametric Instrumental Variable Estimation

There is a vast amount of papers available in the literature on parametric instrumental variables (IV) estimation of econometric models in economics and finance. As with other economic models, one may consider nonparametric structural modeling to permit greater flexibility than tightly specified parametric models in describing such relationships. However, new problems arise for inference in nonparametric structural models that are not present in standard nonparametric regression; see Newey and Powell (2003). Estimation of such models depend on strong regularization and sometimes preclude the asymptotic distribution theory required for inference. To deal with these problems, Newey and Powell (1988) were the first to explore the nonparametric IV models and part of their result was later published in Newey and Powell (2003). Since then, some of the other papers in this area include Newey, Powell and Vella (1999), Daroles, Florens and Renault (2002), Blundell and Powell (2003), Das (2003, 2005), Ai and Chen (2003), Das, Newey and Vella (2003), Newey and Powell (2003), Hall and Horowitz (2005), Cai, Das, Xiong and Wu (2006) (CDXW, hereinafter), Horowitz (2007), and the references therein.

We describe the nonparametric model (with endogenous regressors) below. Suppose we have i.i.d. data $\{(X_i,\, Y_i, Z_i)\}_{i=1}^n$, and the data are generated by the following data generating process:

$$Y_i = g(X_i, Z_{i1}) + u_i, \tag{4.1}$$

where $g(\cdot)$ is an unknown structural function of interest, Z_{i1} is a $d_1 \times 1$ vector of exogenous variables, and the u_i's denote disturbances. The u_i's are correlated with the explanatory variables X_i and, in particular, $E(u_i|X_i) \neq 0$, so that $X_i \in \Re^{d_x}$ is an endogenous variable. Suppose, however, that for each i, we have available another observed data value, $Z_i = (Z_{i1}, Z_{i2})$, for which $E(u_i|Z_i) = 0$, where Z_{i2} is a $d_2 \times 1$ vector of the so-called instrumental variables (IV). Clearly, the nonparametric IV model is different from the standard nonparametric model in the sense that because $E(u_i \,|\, X_i, Z_{i1}) \neq 0$, the structural function $g(\cdot)$ is not given by the regression $E(Y_i \,|\, X_i, Z_{i1})$.

Taking the conditional expectation of (4.1) yields the following integration equation

$$\zeta(z) \equiv E[Y_i|Z_i = z] = E[g(X_i, z_1)|Z_i = z] = \int g(x, z_1) dF_{x|z}(x|z), \tag{4.2}$$

where $F_{x|z}(x|z)$ is the conditional distribution function of X_i given $Z_i = z$. Although $\zeta(z)$ and $F_{x|z}(x|z)$ are estimable based on data $\{(X_i, Y_i, Z_i)\}$, estimation of $g(\cdot)$ is difficult because the relation that identifies $g(\cdot)$ is a Fredholm equation of the first kind, which leads to the difficulty called ill-posed inverse problem in the literature. That is, for nonparametric estimators $\widehat{\zeta}(z)$ and $\widehat{F}_{x|z}(x|z)$ obtained from preliminary nonparametric estimation,

$$\widehat{\zeta}(z) = \int g(x, z_1) d\widehat{F}_{x|z}(x|z)$$

may not exist a solution for $\widehat{g}(\cdot)$. Even if it exists, it may not be computable and continuous in $\widehat{\zeta}(z)$ and $\widehat{F}_{x|z}(x|z)$. As pointed out by Newey and Powell (2003), non-continuity of $\widehat{g}(\cdot)$ is the biggest obstacle to overcome and the lack of continuity of $\widehat{g}(\cdot)$ in $\widehat{\zeta}(\cdot)$ and $\widehat{F}_{x|z}(\cdot)$ means that a small change in $\widehat{\zeta}(\cdot)$ and $\widehat{F}_{x|z}(\cdot)$ may cause a huge error to $\widehat{g}(\cdot)$. Therefore, the consistency of $\widehat{g}(\cdot)$ may not exist even if both $\widehat{\zeta}(\cdot)$ and $\widehat{F}_{x|z}(\cdot)$ are consistent. To recover the structural function $g(\cdot)$ and to overcome these difficulties, in nowadays, several methods were proposed in the literature, described below.

4.1 Series Estimation

Newey and Powell (2003) suggested using the series method to approximate the unknown function $g(\cdot)$ as

$$g(w) \approx \sum_{j=1}^{J} \gamma_j \varphi_j(w), \tag{4.3}$$

where $w = (x, z_1)$, $\{\varphi_j(\cdot)\}$ is a sequence of basis functions and $\{\gamma_j\}$ are the corresponding coefficients. Substitution of (4.3) into (4.2) leads to

$$\zeta(z) = E[Y_i|Z_i = z] \approx \sum_{j=1}^{J} \gamma_j E[\varphi_j(W_i)|Z_i = z] \equiv \sum_{j=1}^{J} \gamma_j p_j(z) = \gamma^T P(z),$$

where $p_j(z) = E[\varphi_j(W_i)|Z_i = z]$, $\gamma = (\gamma_1, \cdots, \gamma_J)^T$ and $P(z) = (p_1(z), \cdots, p_J(z))^T$. Now, to estimate $g(z)$, one can use a nonparametric two-stage approach. At the first stage, using a nonparametric method to obtain $\widehat{p}_j(z)$ and then at the second stage, using the least squares method to obtain $\widehat{\gamma}_j$ by a regression of Y_i on $\{\widehat{p}_j(Z_i)\}$. Finally, one obtains $\widehat{g}(w) = \sum_{j=1}^{J} \widehat{\gamma}_j \varphi_j(w)$. Under some regularity conditions, Newey and Powell (2003) derived the consistency of $\widehat{g}(w)$. But they did not obtain the asymptotic distribution of their estimator.

4.2 Functional Operator Approach

Hall and Horowitz (2005) considered a functional operator approach for estimating $g(\cdot)$. Taking an expectation of $\zeta(Z_i) f_{x,z}(v, Z_i)$ for any fixed v, we have

$$E[\zeta(Z_i) f_{x,z}(v, Z_i)] = \int \zeta(z) f_{z,z}(v, z) f_z(z) dz,$$

where $f_{x,z}(x, z)$ and $f_z(z)$, respectively, denote the joint density of (Z_i, X_i) and the marginal density of Z_i. Substitution of (4.2) into the above equation yields

$$E[\zeta(Z_i) f_{x,z}(v, Z_i)] = \int \int g(x, z_1) f_{x,z}(x, z) f_{z,z}(v, z) dx dz.$$

If one assumes that $g(x, z_1) = g(x)$; that is, $g(\cdot)$ depends only on the endogenous variable X_i but not not on any exogenous variable, then,

$$E[Y_i \, f_{x,z}(v, Z_i)] = E[E(Y_i|Z_i) f_{x,z}(v, Z_i)] = \int g(x) t(x, v) dx \equiv Tg(v),$$

which defines a functional operator T, where

$$t(x,v) = \int f_{x,z}(x,z) f_{z,z}(v,z) dz.$$

Clearly, T is a functional operator defined on the space of functions that are square integrable on $L_2(\Re^{d_x} \times \Re^{d_x})$. Assume that the functional operator T is nonsingular. Then, for each v, $g(v)$ can be expressed as

$$g(v) = E\left[Y_i \left(T^{-1} f_{x,z}\right)(v, Z_i)\right], \tag{4.4}$$

and $g(v)$ could be estimated easily by

$$\widehat{g}(v) = \frac{1}{n} \sum_{i=1}^{n} Y_i \left(T^{-1} f_{x,z}\right)(v, Z_i),$$

if the operator T and $f_{x,z}(v, Z_i)$ were known. Clearly, $f_{x,z}(v, Z_i)$ can be estimated by a kernel method plus jackknife (leave-one-out) approach, given by

$$\widehat{f}_{x,z}(v, Z_i) = \frac{1}{n} \sum_{j=1,\ j \neq i}^{n} K_h(X_j - v, Z_j - Z_i), \tag{4.5}$$

where $K(\cdot, \cdot)$ is a kernel in $\Re^{d_x + d_z}$. Hall and Horowitz (2005) proposed the following estimator

$$\widehat{g}(v) = \frac{1}{n} \sum_{i=1}^{n} Y_i \left(\widehat{T}^{+} \widehat{f}_{x,z}\right)(v, Z_i), \tag{4.6}$$

where $\widehat{T}^{+} = \left(\widehat{T} + a_n I\right)^{-1}$, which is a ridge type estimator and $a_n \to 0$ is a ridge parameter, and

$$\widehat{t}(x,v) = \int \widehat{f}_{x,z}(x,z) \widehat{f}_{z,z}(v,z) dz,$$

where $\widehat{f}_{x,z}(x,z)$ is defined in (4.5). Alternatively, Hall and Horowitz (2005) suggested using a series method to estimate $f_{x,z}(x,z)$; see Hall and Horowitz (2005) for details. Finally, for a general form of $g(x, z_1)$, one can still define the functional operator T_{z_1} for a fixed z_1 and then apply the same idea as above to define the nonparametric estimator for $g(x, z_1)$; see Section 3 of Hall and Horowitz (2005) for the detailed discussions.

Remark 4.1. As addressed in Hall and Horowitz (2005) and Horowitz (2007), equation (4.4) is a Fredholm equation of the first kind. T^{-1} may not always exist and if not, it generates the so-called ill-posed inverse problem. This phenomenon happens if zero is a limit point of the eigenvalues of T, in particular, when $f_{x,z}(x,z)$ is a well behaved density function. In that case, T^{-1} is not a bounded operator, and $g(\cdot)$ cannot be estimated consistently by replacing unknown population quantities on the right-hand side of (4.4) with consistent estimators. This problem is well known in the theory of integral equations. One way to deal with this problem is to modify T^{-1} to make it a continuous operator. Hall and Horowitz (2005) suggested using a ridge idea to

replace T^{-1} for estimation purposes with $(T + a_n I)^{-1}$ (see (4.6) above), where I is the identity operator and $\{a_n\}$ is a sequence of positive constants that converge to 0 as $n \to \infty$.

Hall and Horowitz (2005) derived the asymptotic mean square error of their estimator and showed that for a certain class of distributions, the convergence rates are optimal in a minimax sense, while Horowitz (2007) obtained the asymptotic normality of $\widehat{g}(v)$.

Remark 4.2. For convenience of discussion, assume that $d_x = 1$ (X_i is univariate). Unfortunately, both papers by Hall and Horowitz (2005) and Horowitz (2007) did not discuss whether the convergence rate $(nh)^{-1/2}$ for ordinary nonparametric regression models can be achievable or not, since the convergence rates in both papers depend on the smoothness conditions for the functions $f_{x,z}(\cdot)$ and $g(\cdot)$. To answer the aforementioned question, let us look at Theorem 4.1 of Hall and Horowitz (2005) or Theorem 1 of Horowitz (2007), from which, it follows that the asymptotic integrated mean squared errors (AIMSE) is of the order $O(n^{-(2\beta-1)/(2\beta+\alpha)})$ by using the same notation as in both papers. If it would achieve the optimal convergence rate for ordinary nonparametric regression models, $(2\beta - 1)/(2\beta + \alpha) = 4/5$ so that $\alpha = \beta/2 - 5/4$ which does not satisfy Assumption A3 in Hall and Horowitz (2005) or Assumption 3 in Horowitz (2007). Therefore, one might conclude that the optimal convergence rate for $\widehat{g}(v)$ can not reach the optimal AIMSE rate $O(n^{-4/5})$ for ordinary nonparametric regression models. Finally, both papers mentioned above did not give an explicit expression for the asymptotic bias. Therefore, it is difficult to make the adaptive bandwidth selection feasibly implemented in practice. Now, a natural question arises is whether the optimal convergence rate $(nh)^{-1/2}$ is achievable for a nonparametric estimator under nonparametric IV settings. If possible, it would be interesting to investigate what the scenarios are. Also, it would be warranted to explore the asymptotic bias.

4.3 Projection Method

Newey, Powell and Vella (1999) proposed using a projection method to estimate $g(\cdot)$. The reduced form of (4.1) can be expressed as

$$X_i = \pi(Z_i) + \xi_i, \quad E[\xi_i | Z_i] = 0,$$

where $\pi(Z_i) = E(X_i|Z_i)$. Further, using the new notation $W_i = (\xi_i, X_i, Z_{i1}) \in \Re^{2d_x + d_1}$ and taking the conditional expectation of (4.1) conditional on (X_i, Z_i), we have

$$\begin{aligned} E[Y_i|X_i, Z_i] &= g(X_i, Z_{i1}) + E[u_i|X_i, Z_i] = g(X_i, Z_{i1}) + E[u_i|\xi_i] \\ &\equiv g(X_i, Z_{i1}) + \lambda_0(\xi_i) \equiv h_0(W_i), \end{aligned} \tag{4.7}$$

by assuming that $E[u_i|X_i, Z_i] = E[u_i|\xi_i]$, where the definitions of $\lambda_0(\xi_i)$ and $h_0(W_i)$ should be apparent. Since $E[u_i] = 0$, we have the following projection

$$E[h_0(x, z_1, \xi_i)] = g(x, z_1) + E[\lambda_0(\xi_i)] = g(x, z_1) + E[u_i] = g(x, z_1).$$

Therefore, $g(x, z_1)$ can be estimated by a projection method as

$$\widehat{g}_p(x, z_1) = n^{-1} \sum_{i=1}^{n} \widehat{h}_0(x, z_1, \xi_i),$$

if $\widehat{h}_0(x, z_1, \xi_i)$ and ξ_i would be known. To find a nonparametric estimate $\widehat{h}_0(x, z_1, \xi_i)$ in $\Re^{2d_x+d_1}$, one can use a kernel smoothing technique (say, local linear fitting) as ordinary nonparametric regression by regressing Y_i on $(X_i, Z_{i1}, \widehat{\xi}_i)$, where $\widehat{\xi}_i$ is the nonparametric residual obtained from the reduced form as $\widehat{\xi}_i = X_i - \widehat{\pi}(Z_i)$, where $\widehat{\pi}(Z_i)$ is a nonparametric estimate of $\pi(Z_i)$. Therefore, the feasible estimate $\widehat{g}_p(x, z_1)$ is given by

$$\widehat{g}_p(x, z_1) = \frac{1}{n} \sum_{i=1}^{n} \widehat{h}_0(x, z_1, \widehat{\xi}_i).$$

This method is termed as two-stage nonparametric fitting plus a projection. By following the steps in Masry and Tjøstheim (1997) and Cai and Masry (2000), recently, Su and Ullah (2008) derived the asymptotic properties of the estimator which are the exactly same as that for the ordinary nonparametric regression models. The main disadvantage of using this approach is that it suffers from the problem associated with the curse of dimensionality. Since the unknown function $g(x, z_1)$ is defined in $\Re^{d_x+d_1}$, the nonparametric model fitting has to be implemented in $\Re^{2d_x+d_1}$. This might be infeasible in applications when d_x is large.

Due to the computational convenience and high efficiency in imposing additivity, alternatively, Newey, Powell and Vella (1999) suggested a series method as follows. At the first step, $\pi(Z_i)$ is estimated by

$$\widehat{\pi}(Z_i) = \sum_{j=1}^{K_1} \widehat{\gamma}_j \, r_j(Z_i),$$

where $\{\widehat{\gamma}_j\}$ are obtained by a regression of X_i versus $\{r_j(Z_i)\}$, $\{r_j(Z_i)\}$ is a sequence of basis functions. Then, one obtains the residual $\widehat{\xi}_i = X_i - \widehat{\pi}(Z_i)$. At the second step, a series method is used again as follows. Use the series approximation again to approximate $g(x, z_1)$ and $\lambda_0(\xi)$, respectively, as

$$g(x, z_1) \approx \sum_{l=1}^{K_2} \beta_{l1} \phi_l(x, z_1), \quad \text{and} \quad \lambda_0(\xi) \approx \sum_{m=1}^{K_3} \beta_{m2} \psi_m(\xi),$$

where $\{\phi_l(x, z_1)\}$ and $\{\psi_m(\xi)\}$ are basis functions, so that

$$h_0(w) \approx \sum_{l=1}^{K_2} \beta_{l1} \phi_l(x, z_1) + \sum_{m=1}^{K_3} \beta_{m2} \psi_m(\xi).$$

Then, $\{\beta_{l1}\}$ and $\{\beta_{m2}\}$ can be easily estimated by regressing Y_i versus $\{\phi_l(X_i, Z_{i1})\}$ and $\{\psi_m(\widehat{\xi}_i)\}$. Therefore, $g(x, z_1)$ can be estimated as

$$\widehat{g}_s(x, z_1) = \sum_{l=1}^{K_2} \widehat{\beta}_{l1}\phi_l(x, z_1).$$

Newey, Powell and Vella (1999) derived the consistency of $\widehat{g}_s(x, z_1)$ with a convergence rate for consistency, but they did not derive the asymptotic distribution of their proposed estimator.

4.4 Functional Coefficient Modeling

Das (2005) considered a nonparametric IV model with discrete endogenous variables. That is, X_i is a discrete variable. Without loss of generality, assume that $X_i = 0$ or 1. Then, $g(x, z_1)$ can be rewritten as

$$g(x, z_1) = g(0, z_1)\mathbf{1}(x = 0) + g(1, z_1)\mathbf{1}(x = 1) = a_0(z_1) + a_1(z_1)x,$$

where $a_0(z_1) = g(0, z_1)$ and $a_1(z_1) = g(1, z_1) - g(0, z_1)$. Therefore, $g(x, z_1)$ is linear in endogenous variable but nonlinear in exogenous variable, which is called a functional-coefficient model in the literature; see Cai, Fan and Yao (2000), Li, Huang, Li and Fu (2002), CDXW (2006), Juhl (2005), and Cai and Xu (2008). Assuming that $g(x, z_1)$ has a higher order partial derivative with respect to x, then applying Taylor expansion to $g(x, z_1)$ we obtain

$$g(x, z_1) = \sum_{j=1}^{\infty} \frac{\partial^j g(0, z_1)}{\partial x^j} \frac{x^j}{j!} \approx \sum_{j=0}^{d} a_j(z_1)\, x_j \tag{4.8}$$

for some d, where $a_j(z_1) = \partial^j g(0, z_1)/\partial x^j$ and $x_j = x^j/j!$. This implies that a functional coefficient model might approximate a general nonparametric model well. Therefore, CDXW (2006) studied the following functional coefficient IV model

$$Y_i = \sum_{j=1}^{d} a_j(Z_{i1})^T X_{ij} + u_i = a(Z_{i1})^T X_i + u_i, \quad E[u_i|Z_i] = 0, \tag{4.9}$$

where Y_i is an observable scalar random variable, $\{a_j(\cdot)\}$ are the unknown structural functions of interest, $X_{i0} \equiv 1$, $X_i = (X_{i0}, X_{i1}, \cdots, X_{id})^T$ is a $(d+1)$-dimension vector consisting of d endogenous regressors, $a(Z_{i1}) = (a_0(Z_{i1}), \ldots, a_d(Z_{i1}))^T$, and Z_i is a (d_1+d_2)-dimension vector consisting of a d_1-dimension vector Z_{i1} of exogenous variables and a d_2-dimension vector Z_{i2} of instrumental variables.

Model (4.9) includes the following nonparametric IV model with binary endogenous variable D_i as a special case:

$$Y_i = a_0(Z_{i1}) + a_1(Z_{i1})D_i + \varepsilon_i,$$

which, as noted above, is analyzed in Das (2005). Further, if $a_j(\cdot)$ is a threshold function such as

$$a_j(z) = a_{j1}\,\mathbf{1}(z \le r_j) + a_{j2}\,\mathbf{1}(z > r_j)$$

for some r_j, then model (4.9) may describe a threshold IV regression model. Recently, a threshold model related to this with endogenous covariates has been considered in Caner and Hansen (2004). In this way, the class of models in (4.9) includes some interesting special cases that arise commonly in empirical research.

As elaborated by CDXW (2006), functional coefficient models are appropriate for many applications in economics and finance, and in particular when additive separability of covariates is unsuitable for the problem at hand. For a specific example, CDXW (2006) considered a labor economics problem which is to establish an empirical relationship between marginal returns to education and the level of schooling (see Schultz, 1997). If work experience is also an attribute valued by employers, then the marginal returns to education should vary with experience. As suggested by Card (2001), if a wage model assumes the additive separability of education and experience, the returns to education can be understated at higher levels of education because the marginal return to education is plausibly increasing in work experience. This setting is therefore a natural one for a functional coefficient model, which was further explored by CDXW (2006). Indeed, the marginal returns to education vary positively and nonlinearly with experience and these returns are themselves declining in experience for both low experienced and high experienced workers; see CDXW (2006) for details.

To estimate $\{a_j(z_1)\}$ nonparametrically, CDXW (2006) proposed a two-stage nonparametric method, described as follows. We begin with the first stage, where we obtain $\widehat{\pi}_j(Z_i)$, the fitted value for $\pi_j(Z_i) = E[X_{ij}|Z_i]$ $(1 \le j \le d;\ 1 \le i \le n)$. To this end, we apply the local linear fitting technique and the jackknife (leave-one-out) idea as follows. Assuming that $\{\pi_j(\cdot)\}$ has a continuous second order derivative, when Z_k falls in a neighborhood of Z_i, a Taylor expansion approximates $\pi_j(Z_k)$ by

$$\pi_j(Z_k) \approx \pi_j(Z_i) + (Z_k - Z_i)^T \pi_j'(Z_i) = \alpha_{ij} + (Z_k - Z_i)^T \beta_{ij}.$$

The jackknife idea is to use the all observations except the ith observations in estimating $\pi_j(Z_i)$. Then, the least squares estimator with a local weight (i.e., locally weighted least squares) is given by

$$\sum_{k \ne i}^{n} \left\{ X_{kj} - \alpha_{ij} - (Z_k - Z_i)^T \beta_{ij} \right\}^2 K_{h_1}(Z_k - Z_i).$$

Minimizing the above locally weighted least squares with respect to α_{ij} and β_{ij} gives the local linear estimate of $\pi_j(Z_i)$ by $\widehat{\pi}_{j,-i}(Z_i) = \widehat{\alpha}_{ij}$. Now, we derive the local linear estimator of $\{a_j(\cdot)\}$. The local linear estimators $\widehat{b}_j$ and $\widehat{c}_j$ are defined as the minimizers of the sum of weighted least

squares

$$\sum_{i=1}^{n}\left[Y_i-\sum_{j=0}^{d}\left\{b_j+(Z_{i1}-z_1)^T c_j\right\}\widehat{\pi}_{j,-i}(Z_i)\right]^2 L_{h_2}(Z_{i1}-z_1),$$

and $\widehat{a}_j(z_1)=\widehat{b}_j$, where $L(\cdot)$ is a kernel function at this step.

CDXW (2006) showed that under some regularity conditions,

$$\sqrt{nh_2^{d_1}}\left[\widehat{a}(z_1)-a(z_1)-\frac{h_2^2}{2}\,\mathrm{tr}\left\{\mu_2(L)a''(z_1)\right\}+o_p(h_2^2)\right]\ \xrightarrow{d}\ N(0,\ \Sigma(z_1)), \tag{4.10}$$

where $\Sigma(z_1)=f_{z_1}^{-1}(z_1)\,\nu_0(L)\,\Omega_0^{-1}(z_1)\,\Omega_1(z_1)\,\Omega_0^{-1}(z_1)$, $f_{z_1}(z_1)$ is the marginal density of Z_{i1}, $\Omega_0(z_1)=E[\pi(Z_i)\pi(Z_i)^T|Z_{i1}=z_1]$, and $\Omega_1(z_1)=\Omega_{\eta,1}(z_1)+\Omega_{\xi,1}(z_1)-2\ \Omega_{\eta\xi,1}(z_1)$. The definitions of $\Omega_{\eta,1}(z_1)$, $\Omega_{\xi,1}(z_1)$, and $\Omega_{\eta\xi,1}(z_1)$ can be found in CDXW (2006) and they are omitted here due to too many notations.

One difference of the results in (4.10) compared with those in some other two-stage instrumental regressions (see Newey and Powell, 2003; Newey, Powell and Vella, 1999) is the asymptotic variance term. Here the asymptotic variance consists of three terms: the first addresses the variation of measurement error in the second step, the second term accounts for variability of the estimated reduced form, and the third term accounts correctly for the asymptotic covariance between the first and second steps. The presence of the covariance term is different from some other IV estimators (e.g., Newey, Powell and Vella, 1999), and arises because the second step does not condition on the first step dependent variables.

4.5 Bandwidth Selection

Selecting an optimal (data-driven) bandwidth is an important aspect in applications. Unfortunately, there is basically not an elegant approach to discuss theoretically and empirically how to adaptively select a bandwidth under nonparametric IV settings, when a nonparametric method is applied to estimate the structural regression function, except a rule-of-thumb bandwidth proposed by CDXW (2006) for the functional-coefficient IV models in (4.9). As mentioned in CDXW (2006), the second stage estimation is not sensitive to the choice of the first stage bandwidth so long as the bandwidth h_1 at the first stage is chosen small enough such that the bias in the first stage is not too large. This gives us an ad hoc method to choose h_1, similar to that discussed in Cai (2002a): Use the cross-validation or generalized cross-validation criterion of Cai, Fan and Li (2000) or others to select the bandwidth $\widehat{h}_{10}$, Then use $h_1=A_0\,\widehat{h}_{10}$ ($A_0=1/2$, say, or smaller) or choose a very small h_1 as the first stage bandwidth. Alternatively, A_0 can be taken to be $A_0=n^{-\alpha_1}$ with $\alpha_1>l/(d_1+4)(d_1+l+4)$, as discussed in Cai (2002a), where d_1 is the dimension of the regressor z_1.

In implementation at the second stage, the choice of bandwidth can be carried out as in standard nonparametric regression. In that case, a number of methods could be used to select h_2,

including cross-validation (Stone, 1974), generalized cross-validation (Cai, Fan and Yao, 2000), pre-asymptotic substitution method (Fan and Gijbels, 1996), the plug-in bandwidth selector (Rupert, Sheather and Wand, 1995), empirical bias method (Ruppert, 1997), nonparametric version of the Akaike information criterion (AIC) (see (5.23) later) (Hurvich, Simonoff and Tsai, 1998; Cai and Tiwari, 2000) or the Schwarz-type information criterion (SIC), among others. However, there appears to be no results in the literature for a data-driven bandwidth selection with optimal properties (see Newey, Powell and Vella (1999) for the related discussion) under nonparametric IV settings. It is an open question for future work and it would be very interesting to give a more precise result. Nevertheless, as recommended by CDXW (2006), the procedure suggested above is a useful one for practitioners.

4.6 Semiparametric IV Models

Finally, we would like to mention some recent developments on nonparametric IV models with a parametric part, so that they become semiparametric IV models. Due to the limitation of space, we only cite some references here. First, we mention the paper by Ai and Chen (2003) which discussed a general framework for analyzing economic data (X, Y) by assuming that the data satisfy some conditional moment restrictions such as

$$E[\rho(Z, \theta, m(\cdot))|X] = 0, \tag{4.11}$$

where $Z = (Y^T, X_z^T)^T$, X_z is a subset of X, and $\rho(\cdot)$ is a vector of known (residual) functions. The true conditional distribution of Y given X is assumed unknown and the parameters of interest contain a vector of finite dimensional unknown parameters θ and possibly a vector of infinite dimensional unknown functions $m(\cdot)$. Clearly, if $(Z = (Y_1, Y_2^T, X_1^T, X_2^T)$, $X_z = X_1$ and $\rho(Z_i, \theta, m(\cdot)) = Y_{i1} - \theta^T X_{i1} - m(Y_{i2})$, model (4.11) reduces to a partially linear model

$$Y_{1i} = \beta^T X_{i1} + m(Y_{i2}) + u_i, \tag{4.12}$$

where $E[u_i|X_i] = 0$, which was studied by Newey, Powell and Vella (1999) and Park (2003), while Pakes and Olley (1995) considered a semiparametric IV model with endogenous variables restricted only to the parametric part. Newey, Powell and Vella (1999) used the series method to approximate $m(\cdot)$ and then to estimate both β and $m(\cdot)$ based on the nonparametric series method, whereas Pakes and Olley (1995) and Park (2003) applied the generalized method of moment estimation method to estimate β and $m(\cdot)$.

As argued by Ai and Chen (2003), model (4.11) covers many known nonparametric and semiparametric models as a special case. To estimate θ and $m(\cdot)$, Ai and Chen (2003) proposed to approximate $m(\cdot)$ by a sieve method and then to estimate θ and the sieve parameters jointly by applying the method of minimum distance. They showed that the sieve estimator of

$m(\cdot)$ is consistent with a rate faster than $n^{-1/4}$ under certain metric and the estimator of θ is $\sqrt{n}$-consistent and asymptotically normally distributed. Finally, they addressed the efficiency by choosing the optimally weighted minimum distance to attain the semiparametric efficiency bound. But, they did not provide the asymptotic normality for the sieve estimator of $m(\cdot)$. See Ai and Chen (2003) for details.

To obtain the asymptotic normality of nonparametric part, Cai and Xiong (2006) considered a partially varying coefficient IV model with the following form:

$$Y = g(X,\, Z_1) + \varepsilon = g_1(Z_{11})^T Z_{12} + g_2(Z_{11})^T X_1 + \beta_1^T Z_{13} + \beta_2^T X_2 + \varepsilon, \tag{4.13}$$

where Y is an observable scalar random variable, $X = (X_1^T, X_2^T)^T$ is a vector of endogenous variables including l-dimension vector X_1 and p-dimension vector X_2, $Z_1 = (Z_{11}^T, Z_{12}^T, Z_{13}^T)^T$ is a vector of exogenous variables, consisting of d_{11}-dimension vector Z_{11}, d_{12}-dimension vector Z_{12} with its first element being one, and d_{13}-dimension vector Z_{13}, $Z = (Z_1^T, Z_2^T)^T$ is a d_z-dimension vector with Z_2 being a vector of instrumental variables of dimension d_2, $d_z = d_{11}+d_{12}+d_{13}+d_2$, and $E(\varepsilon \mid Z) = 0$.

To estimate β and $g(\cdot)$ in (4.13), Cai and Xiong (2006) proposed a three-stage method, briefly described below. First, by regarding β as a function of Z_{11}; that is $\beta(Z_{11})$, then model (4.13) becomes (4.9). The nonparametric two-stage proposed in CDXW (2006) can be applied here to estimate $g(\cdot)$ and $\beta(\cdot)$. Note that while β is a global parameter, the estimation of $\beta(\cdot)$ only involves the local data points in a neighborhood of Z_{11} so that the variance is too large. To reduce variance, the estimation of the constant coefficients requires using all data points. Cai and Xiong (2006) proposed using the (weighting) average method to obtain the estimator for β and they showed that the average estimator of β is $\sqrt{n}$-consistent. To address the efficiency of the constant parameter estimator, the weighted version estimator, similar to Ai and Chen (2003), can be used to gain the efficiency by choosing the optimal weighting function to minimize the asymptotic variance. See Cai and Xiong (2006) for the related discussions.

Alternatively, one may use the profile likelihood (least squares for normal likelihood) approach to estimate β_1 and β_2 in (4.13). It is well documented in the literature that for ordinary semiparametric models, profile likelihood is a useful approach and is semiparametrically efficient; see Speckman (1988), Cai (2002a, 2002c), and Fan and Huang (2005) for details. Now we discuss applying the profile likelihood approach to estimate β_1 and β_2 in (4.13). For given β_1 and β_2, model (4.13) becomes

$$Y^* = g_1(Z_{11})^T Z_{12} + g_2(Z_{11})^T X_1 + \varepsilon, \tag{4.14}$$

where $Y^* = Y - \beta_1^T Z_{13} - \beta_2^T X_2$ is the partial residual. This transforms the partially varying coefficient IV model (4.13) into the varying coefficient IV model (4.9). The two-stage local linear estimation technique proposed in CDXW (2006) can be applied to estimate the coefficient

functions $g_1(\cdot)$ and $g_2(\cdot)$, denoted by $\widehat{g}_1(\cdot)$ and $\widehat{g}_2(\cdot)$, respectively. According to CDXW (2006), both $\widehat{g}_1(\cdot)$ and $\widehat{g}_2(\cdot)$ are linear estimators of Y^*. That is,

$$\widehat{\mathbf{M}} = \begin{pmatrix} \widehat{g}_1(Z_{11,1})^T Z_{12,1} + \widehat{g}_2(Z_{11,1})^T X_{1,1} \\ \vdots \\ \widehat{g}_1(Z_{11,n})^T Z_{12,n} + \widehat{g}_2(Z_{11,n})^T X_{1,n} \end{pmatrix} = \mathbf{S}\,\mathbf{Y}^* = \mathbf{S}\,(\mathbf{Y} - \mathbf{Z}_{13}\,\beta_1 - \mathbf{X}_2\,\beta_2),$$

where $\mathbf{Y}^* = (Y_1^*, \cdots, Y_n^*)^T$. The matrix $\mathbf{S}$ is a smoothing matrix and depends only on the data $\{(Z_{11,i}, Z_{12,i}, X_{1,i}, \widehat{X}_{1,i}), i = 1, \cdots, n\}$ and the kernel function, where $\widehat{X}_{1,i}$ is obtained from the reduced equation by the jackknife least squares method; see CDXW (2006) for the explicit expression for $\mathbf{S}$ and $\widehat{X}_{1,i}$ (which depends on the data $\{(X_j, Z_j), j = 1, \cdots, i-1, i+1, \cdots, n\}$). Substituting $\widehat{\mathbf{M}}$ into (4.14), we obtain the following linear IV model

$$(\mathbf{I} - \mathbf{S})\mathbf{Y} = (\mathbf{I} - \mathbf{S})[\mathbf{Z}_{13}\,\beta_1 + \mathbf{X}_2\,\beta_2] + \boldsymbol{\varepsilon}. \tag{4.15}$$

Applying the two-stage least squares to the linear model (4.15), we obtain the profile likelihood estimators of β_1 and β_2, respectively, termed as *profile two-stage least squares estimate*. Note that if there is no endogeneity in the model, Fan and Huang (2005) showed that the profile likelihood estimator is semiparametrically efficient. Therefore, we conjecture that the profile least squares estimate for β_2 described above should be $\sqrt{n}$-consistent and semiparametrically efficient. It is interesting to justify this result theoretically.

5 Nonparametric Quantile Regression Models

Since quantile regression or conditional quantile was introduced by Koenker and Bassett (1978), it has been successfully and widely used in various disciplines, such as finance, economics, medicine, and biology. In nowadays, estimation of conditional quantiles is a common practice in risk management operations and many other financial applications. The literature on estimating quantile regression function is large but is still swiftly growing. Much of the study on quantile regression is based on linear parametric quantile regression models. But in recent years, nonparametric quantile regression models in both theory and applications have attracted a great deal of research attentions due to their greater flexibility than tightly specified parametric models. A non-exhaustive list of important recent contributions to this growing literature include (but not limited to) Chaudhuri (1991), Koenker, Portnoy and Ng (1992), Fan, Hu and Troung (1994), Koenker, Ng and Portnoy (1994), Chaudhuri, Doksum and Samarov (1997), He, Ng and Portnoy (1998), Yu and Jones (1998), He and Ng (1999), He and Portnoy (2000), Honda (2000, 2004), Khindanova and Rachev (2000), Cai (2002b), Cai and Ould-Said (2003), De Gooijer and Zerom (2003), Yu and Lu (2004), Engle and Manganelli (2004), Horowitz and Lee (2005), Kim (2007), and Cai and Xu (2008) and references therein for recent statistics and econometrics literature on nonparametric estimation of quantile regression models.

Let $\{X_t, Y_t\}_{t=1}^n$ be a stationary sequence and $F(y\,|x)$ denote the conditional distribution of Y_t given $X_t = x$, where X_t is a vector of covariates in $\Re^d$, including possibly exogenous variables and lagged variables, the conditional quantile function of Y_t given $X_t = x$ is defined as, for any $0 < \tau < 1$,

$$q_\tau(x) = \inf\{y \in \Re : F(y\,|x) \geq \tau\} = F^{-1}(\tau\,|x), \tag{5.1}$$

where $F^{-1}(\tau\,|x)$ is the inverse function of $F(y\,|x)$. Equivalently, $q_\tau(x)$ can be expressed as

$$q_\tau(x) = \text{argmin}_{a\in\Re} E\left\{\rho_\tau(Y_t - a)\,|\,X_t = x\right\}, \tag{5.2}$$

where $\rho_\tau(y) = y\,[\tau - I\{y < 0\}]$ with $y \in \Re$ is called the loss ("check") function and $I\{A\}$ is the indicator function of any set A. Function $q_\tau(x)$ is called as a conditional quantile function or regression quantile.

It is well documented that quantile regression has several important properties, described as follows. It does not require knowing the distribution of Y_t and symmetry of the distribution. When $\tau = 1/2$, it becomes the median or least absolute deviation regression which is well known to posses the robustness. Therefore, it has a robust property. Also, it has an ability to model heterogeneous effects and to account for unobserved heterogeneity. To see the intuitive behind this property, we use the basic Skorohod representation to express the quantile regression model. Using this representation, the dependent variable Y_t, conditional on the exogenous variable of interest X_t, takes the form

$$Y_t = q(X_t, U_t), \quad \text{where} \quad U_t|X_t \sim U(0,1),$$

where $q(x,u) = q_u(x)$ is the conditional u-th quantile of Y_t given $X_t = x$ and U_t is the nonseparable error. Furthermore, it is convenient to use the conditional quantile for detecting conditional heteroskedasticity. To this end, we assume that Y_t is related to X_t through the model

$$Y_t = m(X_t) + \sigma(X_t)\,\varepsilon_t,$$

where $m(\cdot)$ is the mean function, $\sigma^2(\cdot)$ is the variance function, and X_t and ε_t are independent. The conditional quantile of Y_t given X_t is

$$q_\tau(X_t) = m(X_t) + \sigma(X_t)\,F_{\varepsilon_t}^{-1}(\tau),$$

where $F_{\varepsilon_t}(\cdot)$ is the distribution of ε_t. An informal way to test conditional heteroskedasticity is to use a graph. That is, if the curves of $q_\tau(x)$ for different values of τ are parallel, this indicates that $\sigma(\cdot)$ should be a constant. Moreover, regression quantiles can also be useful for the estimation of predictive intervals. For example, in predicting the response from a given covariate X_t, estimates of $q_{\alpha/2}(X_t)$ and $q_{1-\alpha/2}(X_t)$ can be used to obtain a $(1-\alpha)\,100\%$ nonparametric predictive interval. Finally, it is very useful in various applied fields. For example, in risk

management, it can be used to compute the conditional value-at-risk (CVaR): the percentage loss in market value over a given time horizon that is exceeded with a certain probability, and the conditional expected shortfall (CES). Indeed, CVaR can be regarded as a special case of quantile regression. Of course, there are many methods available to model the CVaR. The conditional expected shortfall can be expressed in terms of a regression quantile as

$$E[Y_t|Y_t \leq q_\tau(X_t), X_t] = \int_0^\tau q_u(X_t)du/\tau.$$

For details, see Cai and Wang (2008).

Given observed data $\{X_t, Y_t\}_{t=1}^n$, the main interest is to estimate $q_\tau(x)$. If we assume that $q_\tau(x) = \beta_\tau^T x$, we obtain a linear quantile regression model, which is popular in the literature; see the book by Koenker (2005), and we can estimate easily the parameters (see (5.17) below). In some practical applications, a linear quantile regression model might not be flexible enough to capture the underlying complex dependence structure. For example, some components may be highly nonlinear or some covariates may be interactive. Therefore, to make quantile regression models more flexible, there is a swiftly growing literature on nonparametric quantile regression. Various smoothing techniques, such as kernel methods, splines, and their variants, have been used to estimate the nonparametric quantile regression for both independent and time series data. Some recent developments and detailed discussions on theory, methodologies, and applications can be found in the literature. For example, Chaudhuri (1991), Fan, Hu and Troung (1994), Chaudhuri, Doksum and Samarov (1997), Yu and Jones (1998), Honda (2000), Cai (2002b), and Cai and Ould-Said (2003) considered nonparametric kernel smoothing estimate of quantile function, while He, Ng and Portnoy (1998), He and Ng (1999), and He and Portnoy (2000) used spline methods to obtain nonparametric estimate. However, a purely nonparametric quantile regression model may suffer from the so-called "curse of dimensionality" problem, the practical implementation might not be easy, and the visual display may not be useful for the exploratory purposes. To deal with the aforementioned problems, some dimension reduction modelling methods have been proposed in the literature. For example, De Gooijer and Zerom (2003), Yu and Lu (2004), and Horowitz and Lee (2005) considered the additive quantile regression models for iid data, while Honda (2004) and Cai and Xu (2008) investigated the varying coefficient quantile regression models for time series processes. Particularly, there has been some study on a time-varying coefficient quantile regression model, which is potentially useful to see whether the quantile regression changes over time and in a case with a practical interest is, for example, the analysis of the reference growth data by Cole (1994), Wei, Pere, Koenker and He (2006), Wei and He (2006), and Kim (2007).

5.1 Direct Methods

A direct procedure is based on equation (5.1), described as follows. First, estimate the conditional distribution function using a nonparametric method such as the "double-kernel" local linear technique (LL) of Yu and Jones (1998) and then to invert the conditional distribution estimator to produce an estimator of a conditional quantile. This estimator is called the Yu and Jones estimator (see $\widehat{q}_{\tau,ll}(x)$ in (5.14) later); see Yu and Jones (1998) for details. As noticed by Cai (2002b) and Cai and Wang (2008), the key for a direct estimation method is to find a good estimator for conditional distribution function. Further, as demonstrated by Cai (2002b), although local linear estimators of the Yu and Jones type have some attractive properties such as no boundary effects, design adaptation, and mathematical efficiency; see, e.g., Fan and Gijbels (1996), they have the disadvantage of producing conditional distribution function estimators that are not constrained either to lie between zero and one or to be monotone increasing although some modifications in implementation were addressed by Yu and Jones (1998). In both these respects, the Nadaraya-Watson (NW) methods are superior, despite their rather large bias and boundary effects. The properties of positivity and monotonicity are particularly advantageous if the method of inverting the conditional distribution estimator is applied to produce an estimator of a conditional quantile.

To overcome these difficulties, Cai (2002b) and Cai and Wang (2008) proposed a weighted version of the NW (WNW) estimator and weighted double kernel estimator (WDK), which are designed to possess the superior properties of local linear methods such as bias reduction and no boundary effect and to preserve the property that the NW estimator is always a distribution function. Cai (2002b) and Cai and Wang (2008) established the asymptotic normality and weak consistency for both the WNW and WDK estimators of conditional distribution for α-mixing under a set of weaker conditions at both boundary and interior points. It is therefore shown, to the first order, that the WNW method enjoys the same convergence rates as those of the local linear "double-kernel" procedure of Yu and Jones (1998). More importantly, both the WNW and WDK estimators have desired sampling properties at both boundary and interior points of the support of the design density. Cai (2002b) and Cai and Wang (2008) also derived both the WNW and WDK estimators of the conditional quantile by inverting their estimated conditional distributions estimator and showed that both the WNW and WDK quantile estimators always exist as a result of both the WNW and WDK distributions being a distribution function in finite samples and that they inherit all advantages from the WNW and WDK estimators of conditional distribution.

For simplicity of notation, we consider the case of $d = 1$. We now turn to the estimation of the conditional distribution function $F(y|x)$. To this end, let $p_t(x)$, for $1 \leq t \leq n$, denote the weight functions of the data $X_1, \ldots, X_n$ and the design point x with the property that each

$p_t(x) \geq 0$, $\sum_{t=1}^n p_t(x) = 1$ and

$$\sum_{t=1}^{n} (X_t - x)\, p_t(x)\, K_h(x - X_t) = 0, \tag{5.3}$$

where $K(\cdot)$ is a kernel function, $K_h(\cdot) = K(\cdot/h)/h$, and $h = h_n > 0$ is the bandwidth. Motivated by the property of local linear estimator, the constraint (5.3) can be regarded as a discrete moment condition; see Fan and Gijbels (1996, p.63) for details. Of course, $\{p_t(x)\}$ satisfying these conditions are not uniquely defined and we specify them by maximizing $\prod_{t=1}^n p_t(x)$ subject to the constraints. The weighted version of Nadaraya-Watson estimator of the conditional distribution $F(y \,|\, x)$ of Y_t given $X_t = x$ is defined

$$\widehat{F}_{wnw}(y \,|\, x) = \frac{\sum_{t=1}^n p_t(x)\, K_h(x - X_t)\, \mathbf{1}(Y_t \leq y)}{\sum_{t=1}^n p_t(x)\, K_h(x - X_t)}.$$

Note that $0 \leq \widehat{F}_{wnw}(y \,|\, x) \leq 1$ and it is monotone in y. Cai (2002b) showed that $\widehat{F}_{wnw}(y \,|\, x)$ is first-order equivalent to a local linear estimator (see $\widehat{F}_{ll}(y \,|\, x)$ in (5.13) later). More importantly, that $\widehat{F}_{wnw}(y \,|\, x)$ has automatic good behavior at boundaries. In contrast, $\widehat{F}_{ll}(y \,|\, x)$ may not take values in [0,1] and it may not be monotone in y.

The natural question arises regarding how to choose the weights. Borrowing the idea is from the empirical likelihood, Cai (2002b) suggested maximizing $\sum_{t=1}^n \log\{p_t(x)\}$ subject to the constraints $\sum_{t=1}^n p_t(x) = 1$ and (5.3) through the Lagrange multiplier method, the $\{p_t(x)\}$ are simplified to

$$p_t(x) = n^{-1}\, \{1 + \lambda\, (X_t - x)\, K_h(x - X_t)\}^{-1},$$

where λ, a function of data and x, is uniquely defined by (5.3), which ensues that $\sum_{t=1}^n p_t(x) = 1$. Equivalently, λ is chosen to maximize

$$L_n(\lambda) = \frac{1}{n\, h} \sum_{t=1}^{n} \log\{1 + \lambda\, (X_t - x)\, K_h(x - X_t)\}. \tag{5.4}$$

In implementation, Cai (2002b) recommended using the Newton Raphson scheme to find the root of equation $L_n'(\lambda) = 0$.

Cai (2002b) showed that, under some regularity conditions including that $\{(X_t, Y_t)\}_{t=1}^n$ is an α-mixing sequence, then as $n \to \infty$,

$$\widehat{F}_{wnw}(y \,|\, x) - F(y \,|\, x) = \frac{1}{2}\, h^2\, \mu_2(K)\, F^{2,0}(y \,|\, x) + o_p(h^2) + O_p\left((n\, h)^{-1/2}\right), \tag{5.5}$$

where $F^{a,b}(y \,|\, x) = \partial^{a+b}/\partial y^a \partial x^b F(y \,|\, x)$ and $\mu_j(K) = \int u^j K(u) du$. This, of course, implies that $\widehat{F}_{wnw}(y \,|\, x) \to F(y \,|\, x)$ in probability with a rate. In addition, Cai (2002b) derived the asymptotic normality for $\widehat{F}_{wnw}(y \,|\, x)$ as

$$\sqrt{n\, h}\, \left[\widehat{F}_{wnw}(y \,|\, x) - F(y \,|\, x) - B_f(y \,|\, x) + o_p(h^2)\right] \xrightarrow{d} N\left(0,\, \sigma_f^2(y \,|\, x)\right), \tag{5.6}$$

where the bias and variance are given respectively by

$$B_f(y\,|\,x) = \frac{1}{2}\,h^2\,\mu_2(K)\,F^{2,0}(y\,|\,x), \quad \text{and} \quad \sigma_f^2(y\,|\,x) = \nu_0(K)\,F(y\,|\,x)[1-F(y\,|\,x)]/f_1(x) \quad (5.7)$$

with $f_1(x)$ being the marginal density of X_t. This implies that to the first order, the WNW method enjoys the exactly same convergence rates as those of local linear "double-kernel" procedure (see $\widehat{F}_{ll}(y\,|\,x)$ in (5.13) later) of Yu and Jones (1998), under similar regularity conditions. However, Yu and Jones (1998) treated only the case of independent data.

Based on (5.1), we define the WNW type conditional quantile estimator $\widehat{q}_{wnw}(x)$ to satisfy $\widehat{F}_{wnw}\left(\widehat{q}_{wnw}(x)\,|\,x\right) = \tau$ so that

$$\widehat{q}_{wnw}(x) = \inf\left\{y \in \Re : \widehat{F}_{wnw}(y\,|\,x) \geq \tau\right\} \equiv \widehat{F}^{-1}_{wnw}(\tau\,|\,x). \quad (5.8)$$

Clearly, $\widehat{q}_{wnw}(x)$ always exists since $\widehat{F}_{wnw}(y\,|\,x)$ is between 0 and 1 and monotone in y, and it involves only one bandwidth so that it makes practical implementation more appealing. In contrast, the local linear double-kernel estimator of Yu and Jones (1998) has some difficulty of inverting the conditional distribution estimator due to lack of monotonicity and it requires choosing two bandwidths although the second bandwidth should not be very sensitive (see Remark 5.1 later). Furthermore, Cai (2002b) showed that the WNW estimator $\widehat{q}_{\tau,wnw}(x)$ maintains the aforementioned advantages as $\widehat{F}_{wnw}(y|x)$ does. Also, Cai (2002b) showed that under some regularity conditions, as $n \to \infty$,

$$\sqrt{n\,h}\,\left[\widehat{q}_{\tau,wnw}(x) - q_\tau(x) - B_\tau(x) + o_p(h^2)\right] \;\xrightarrow{d}\; N\left(0,\,\sigma_\tau^2(x)\right), \quad (5.9)$$

where the bias and variance are given respectively by

$$B_\tau(x) = -\frac{B_f(q_\tau(x)\,|\,x)}{f(q_\tau(x)\,|\,x)} \quad \text{and} \quad \sigma_\tau^2(x) = \frac{\sigma_f^2(q_\tau(x)\,|\,x)}{f^2(q_\tau(x)\,|\,x)} = \frac{\nu_0(K)\,p[1-p]}{f^2(q_\tau(x)\,|\,x)\,f_1(x)}, \quad (5.10)$$

where $f(y|x)$ is the conditional density of $Y_t = y$ given $X_t = x$.

It is clear that for given x, $\widehat{F}_{wnw}(y\,|\,x)$ is not a continuous function of y. It might cause the computational trouble when computing the estimated conditional quantile $\widehat{q}_{\tau,wnw}(x)$ by (5.8). To overcome this shortcoming, Cai and Wang (2008) proposed a weighted double kernel estimator (see below), which indeed is differentiable with respect to y. Cai and Wang (2008) showed that the differentiability of the estimated conditional distribution function can not only make the asymptotic analysis much easier for the nonparametric estimators of quantile regression, but also can reduce the asymptotic variance (or asymptotic mean squared error) in a higher order sense. The main idea of Cai and Wang (2008) is described as follows.

It is noted for a given symmetric kernel $g(\cdot)$ where $G(\cdot)$ is the distribution function of $g(\cdot)$, as $h_0 \to 0$,

$$E\{G_{h_0}(y - Y_t)\,|\,X_t = x\} = F(y\,|\,x) + \frac{h_0^2}{2}\,\mu_2(g)\,F^{0,2}(y\,|\,x) + o(h_0^2) \to F(y\,|\,x), \quad (5.11)$$

where $G_{h_0}(u) = G(u/h_0)/h_0$. The above convergence ignores the higher terms $o(h_0^2)$ since $h_0 = o(h)$, where h is the smoothing bandwidth in the x direction (see (5.12) below). We can see that $Y_t^*(y) = G_{h_0}(y - Y_t)$ can be regarded as an initial estimate of $F(y \,|\, x)$ smoothing in the y direction. Thus, the left hand side of (5.11) can be regraded as a nonparametric regression of the observed variable $Y_t^*(y)$ versus X_t and the local linear (or polynomial) fitting scheme can be applied here. This leads to the locally weighted least squares regression problem:

$$\sum_{t=1}^{n} \{Y_t^*(y) - a - b\,(X_t - x)\}^2 \; K_h(x - X_t). \tag{5.12}$$

Note that (5.12) involves two kernels $g(\cdot)$ and $K(\cdot)$ and two bandwidths h_0 and h. This is the reason for calling it "double kernel".

Minimizing (5.12) with respect to a and b, we obtain the locally weighted least squares estimator of $F(y \,|\, x)$, which is $\widehat{a}$. It is easy to see that this estimator can be re-expressed as a linear estimator as

$$\widehat{F}_{ll}(y \,|\, x) = \sum_{t=1}^{n} W_{ll,t}(x, h)\, G_{h_0}(y - Y_t), \tag{5.13}$$

where with $S_{n,j}(x) = \sum_{t=1}^{n} K_h(x - X_t)\,(X_t - x)^j$, the weights $\{W_{ll,t}(x, h)\}$ are given by

$$W_{ll,t}(x, h) = [S_{n,2}(x) - (x - X_t)\, S_{n,1}(x)]\, K_h(x - X_t)\, \left[S_{n,0}(x) S_{n,2}(x) - S_{n,1}^2(x)\right]^{-1}.$$

Clearly, $\{W_{ll,t}(x, h)\}$ satisfy the discrete moments conditions given (5.3). $\widehat{F}_{ll}(y \,|\, x)$ is the so-called Yu and Jones estimator. Yu and Jones (1998) studied the asymptotic properties of $\widehat{F}_{ll}(y \,|\, x)$ for iid data, which are similar to those given in (5.5) and (5.6) if $h_0 = o(h)$.

Remark 5.1. If the bandwidth at the initial step h_0 is not under-smoothed, say $h_0 = O(h)$, then there is an extra term in the asymptotic bias and it is given by $\mu_2(g)\,(h_0^2/2)\, F^{0,2}(y \,|\, x)$, which is carried over from the initial estimation.

Also, Yu and Jones (1998) considered the nonparametric estimate of $q_\tau(x)$ based on $\widehat{F}_{ll}(y|x)$, which is defined as

$$\widehat{q}_{\tau,ll}(x) = \widehat{F}_{ll}^{-1}(\tau|x), \tag{5.14}$$

and they derived the asymptotic properties of $\widehat{q}_{\tau,ll}(x)$, which is the exactly same as that given in (5.9). Further, Yu and Jones (1998) proposed an ad hoc method to adaptively select the optimal bandwidths h_0 and h. Clearly, $\widehat{F}_{ll}(y \,|\, x)$ may not be constrained either to lie between zero and one or monotone increasing. To overcome this difficulty, some modifications in implementation of $\widehat{q}_{\tau,ll}(x)$ were addressed in Yu and Jones (1998).

To accommodate all of the above attractive properties (monotonicity, continuity, differentiability, lying between zero and one, design adaption, avoiding boundary effects, and mathematical efficiency) of both estimators $\widehat{F}_{ll}(y \,|\, x)$ and $\widehat{F}_{wnw}(y \,|\, x)$ under a unified framework, Cai

and Wang (2008) proposed the following nonparametric estimator for conditional distribution $F(y\,|\,x)$, termed as weighted double kernel estimation,

$$\widehat{F}_{wdk}(y\,|\,x) = \sum_{t=1}^{n} W_{wdk,t}(x,h)\, G_{h_0}(y - Y_t), \tag{5.15}$$

where

$$W_{wdk,t}(x,h) = p_t(x)\, W_h(x - X_t) \left[\sum_{t=1}^{n} p_t(x)\, W_h(x - X_t)\right]^{-1},$$

and $\{p_t(x)\}$ is chosen to be $p_t(x) = n^{-1}\,\{1 + \lambda\,(X_t - x)\, W_h(x - X_t)\}^{-1} \geq 0$ to satisfy (5.3). Here λ is a function of the data and x and is uniquely defined by (5.4). Cai and Wang (2008) showed that the asymptotic properties for $\widehat{F}_{wdk}(y\,|\,x)$ are similar to those given in (5.5) and (5.6) if $h_0 = o(h)$. Note that this under-smoothing at the initial step is needed (see Remark 5.1).

Moreover, Cai and Wang (2008) considered the nonparametric estimate of $q_\tau(x)$ based on $\widehat{F}_{wdk}(y|x)$, which is defined as

$$\widehat{q}_{\tau,wdk}(x) = \widehat{F}^{-1}_{wdk}(\tau|x). \tag{5.16}$$

Note that $\widehat{q}_{\tau,wdk}(x)$ always exists in finite samples and is uniquely determined since $\widehat{F}_{wdk}(y|x)$ is a continuous distribution function. Cai and Wang (2008) also showed that $\widehat{q}_{\tau,wdk}(x)$ has the exactly same asymptotic behavior as that given in (5.9). In addition, Cai and Wang (2008) proposed an ad hoc data-driven bandwidth selection method based on the nonparametric version of the Akaike information criterion.

Finally, Yu and Jones (1998), Cai (2002b) and Cai and Wang (2008) discussed the asymptotic behavior of their nonparametric estimators $\widehat{q}_{\tau,ll}(x)$, $\widehat{q}_{\tau,wnw}(x)$ and $\widehat{q}_{\tau,wdk}(x)$ at boundaries and the result shows that all estimators have the exactly same asymptotic bias and do not have boundary effect; see Yu and Jones (1998), Cai (2002b) and Cai and Wang (2008) for details.

Cai and Wang (2008) considered a real data set on Dow Jones Industrials (DJI) index returns and applied the proposed method to estimate the 5% CVaR and CES functions. Both the CVaR and CES estimates exhibit a U-shape, which corresponds to the so-called "volatility smile". Therefore, the risk tends to be lower when the lagged log loss of DJI is close to the empirical average, and larger otherwise. We can also observe the curves are asymmetric. This may indicate that the DJI index is more likely to fall if there were a loss within the last day than if there was a same amount of positive return.

5.2 Loss Function Approaches

Based on (5.2), if $q_\tau(x) = \beta_\tau^T x$ is linear in x, then, one can find the estimate of β_τ by

$$\widehat{\beta}_\tau = \text{argmin}_{\beta_\tau} \sum_{t=1}^{n} \rho_\tau(Y_t - \beta_\tau^T x); \tag{5.17}$$

see Koenker and Bassett (1978, 1982) for details.

To compute $\widehat{\beta}_\tau$ in (5.17), it can be implemented by using the function **rq()** in the package **quantreg** in the computing language **R**, due to Koenker (2004).

If $q_\tau(x)$ is a nonparametric function, there are several methods proposed in the literature to estimate $q_\tau(x)$, we describe some of them below.

5.2.1 Local Polynomial Methods

If $q_\tau(x)$ is assumed to have continuous $(m+1)$th order partial derivative, for X_t in a neighborhood of x, $q_\tau(X_t)$ can be approximated by $\sum_{j=0}^m \theta_j(X_t - x)^j$, where $\theta_j = (1/j!)\partial^j q_\tau(x)/\partial x^j$ is the jth partial derivative of $q_\tau(x)$. Then, we can use the following locally weighted loss function, which is a locally weighted version of (5.17),

$$\widehat{\theta} = \text{argmin}_\theta \sum_{t=1}^n \rho_\tau \left(Y_t - \sum_{j=0}^m \theta_j (X_t - x)^j \right) K_h(x - X_t) \tag{5.18}$$

to obtain the local polynomial estimation of quantile function. Clearly, $\widehat{q}_\tau(x) = \widehat{\theta}_0$ estimates the quantile function and $\widehat{q}_\tau^{(j)}(x) = j!\,\widehat{\theta}_j$ estimates the jth partial derivative. Note that formula (5.18) has been addressed (essentially) by Chaudhuri (1991), Fan, Hu and Troung (1994), Koenker, Portnoy and Ng (1992), Yu and Jones (1998) for iid sample and Honda (2000) and Cai and Ould-Said (2003) for time series.

To compute $\widehat{q}_\tau(x)$ and $\widehat{q}_\tau^{(j)}(x)$, one also can use the function **rq()** by setting covariates as $X_t - x$, $\cdots$, $(X_t - x)^m$, and the weight as $K_h(X_t - x)$. Alternatively, one can use the function **lprq()** in the same package.

By using the series expansion method, Chaudhuri (1991) was the first to obtain the local Bahadur type representation of parameter's estimators so that one can easily derive some asymptotic results. Honda (2000) generalized these results to the α-mixing process by using local polynomial fitting, and obtained the similar asymptotic results. To derive the asymptotic properties, Honda (2000) and Cai and Xu (2008) gave the local Bahadur representation for $\widehat{q}_\tau(x)$ for univariate case ($d = 1$). That is, they showed that under some regular conditions, the local linear ($m = 1$) quantile estimator $\widehat{q}_\tau(x)$ has the following representation,

$$\sqrt{nh}\,[\widehat{q}_\tau(x) - q_\tau(x)] = \frac{1}{f_{y|x}(q_\tau(x)|x) f_1(x) \sqrt{nh}} \sum_{t=1}^n \psi_\tau(Y_t^*) K((X_t - x)/h) + o_p(1), \tag{5.19}$$

where $\psi_\tau(x) = \tau - I_{x<0}$ and $Y_t^* = Y_t - q_\tau(x) - q'_\tau(x)(X_t - x_0)$. Therefore, one can easily obtain the asymptotic normality as

$$\sqrt{nh} \left[\widehat{q}_\tau(x) - q_\tau(x) - \frac{h^2}{2} \mu_2(K)\, q''_\tau(x) + o_p(h^2) \right] \xrightarrow{d} N(0,\, \sigma_\tau^2(x)), \tag{5.20}$$

where $\sigma_\tau^2(x)$ is given in (5.10). Clearly, a comparison of (5.9) and (5.20) leads to conclusions that the local linear quantile estimator $\widehat{q}_\tau(x)$ and three direct estimators share the exactly same asymptotic variance, but the biases are quite different. Indeed, the bias term in (5.9) (see also (5.10)), the quantity $-F^{2,0}(q_\tau(x)\,|\,x)/f(q_\tau(x)\,|\,x)$, involving the second derivative of the conditional distribution function, is replaced by $q''_\tau(x)$, the second derivative of the conditional quantile function itself. This is not surprising since for the direct methods, the approximation is applied to the conditional distribution function, while for local linear quantile estimator $\widehat{q}_\tau(x)$, the approximation is applied to the conditional quantile function itself.

5.2.2 Spline Approaches

In the 1990's, there were many research papers on nonparametric estimation of quantile regression using various splines methods such as smoothing splines and B-splines. For example, for a single covariate, He and Shi (1994) used quantile regression B-splines and considered the convergence with a rate of B-splines estimator, while Koenker, Ng and Portnoy (1994) suggested quantile smoothing splines. In bivariate smoothing, He, Ng and Portnoy (1998) considered bivariate quantile smoothing splines that belong to the space of bilinear tensor product splines, while Portnoy (1997) and He and Portnoy (2000) provided the asymptotic properties of these bivariate quantile splines estimators. The optimality properties of the splines provide justification for their use in nonparametric quantile function estimation, and the optimization problems can be solved efficiently as linear programs. He and Ng (1999) considered a general additive (several covariates) model with univariate linear splines capturing the main effects and bilinear tensor product splines capturing the second order interactions. But all splines methods encounter the same difficulties that it is not easy to derive the asymptotic properties like asymptotic normality and to make statistical inferences (see Remark 5.3 later for more discussions), although they might be attractive in applications.

We now begin by briefly reviewing the smoothing splines technique; see the aforementioned papers for details. For a univariate design variable X_t with observed response Y_t, the τ-th quantile smoothing spline function $q_\tau(x)$ minimizes over

$$\sum_{t=1}^{n} \rho_\tau\left(Y_t - q_\tau(X_t)\right) + \lambda\, V(q'_\tau), \tag{5.21}$$

where $V(h) = \sup \sum_{j=1}^{k} |h(x_j) - h(x_{j-1})|$ denotes the total variation of the function $h(\cdot)$ with the supremum being taken over all finite partitions $x_0 < x_1 < \cdots < x_k$ of the support of $h(\cdot)$. If $h(\cdot)$ is differentiable, it is easy to see that

$$V(h) = \int_0^1 |h'(x)|dx, \quad \text{if the support of } h(\cdot) \text{ is } [0,\ 1].$$

The optimal solution $\widehat{q}_\tau(x)$ estimates the τ-th conditional quantile function $q_\tau(x)$. The problem of quantile smoothing in expression (5.21) can be viewed as a special case ($p = 1$) of the following general form of quantile smoothing

$$\sum_{t=1}^{n} \rho_\tau \left(Y_t - q_\tau(X_t) \right) + \lambda \left(\int |q''_\tau(x)|^p \, dx \right)^{1/p} \tag{5.22}$$

for $p \geq 1$. If $p = 2$ in (5.22), the solution to expression (5.22) is a natural cubic smoothing spline with knots at the observed design points. Its computation is rather efficient as it simply amounts to solving a linear system. The solution to expression (5.21) is a linear smoothing spline with possible breaks in the derivative at the design points, and the computation can be performed by modern linear programming methods. See the forgoing papers for the computational issue. As for selecting the smoothing parameter λ, the Schwarz-type information criterion is commonly suggested in the smoothing spline literature; see Koenker, Ng and Portnoy (1994) and He and Ng (1999) for details. But it is well known that the SIC is over-fitting due to the heavy penalty (see (5.23) later) when the sample is large.

Remark 5.2. As commented by He, Ng and Portnoy (1998), generalization of smoothing splines to bivariate or multivariate cases is not always straightforward. The form of the solution often depends on the roughness penalty used in the optimization process and it is quite complex. Due to the complicated notation, we ignore the presentation of smoothing splines for multivariate case. Instead, we refer the reader to the papers by He and Shi (1994), He, Ng and Portnoy (1998), He and Ng (1999), and He and Portnoy (2000) for the detailed discussions.

Remark 5.3. It is well known in the splines literature; see the previously mentioned papers, that the rate of convergence for the nonparametric estimates depends mainly on two aspects: the smoothness of the function being estimated and the dimensionality of the spline space or, equivalently, the number of knots. These issues are still valid for the conditional quantile smoothing splines estimates. The asymptotic behavior such as the rate of convergence for the quantile smoothing splines is rather difficult to analyze, especially when a data-driven smoothing parameter is used. In the univariate case when the smoothing parameter is not data-driven, Portnoy (1997) derived some local asymptotic properties of the quantile smoothing splines, while He and Ng (1999) and He and Portnoy (2000) presented the asymptotic mean square error for bivariate and multivariate cases. Unfortunately, the asymptotic normality of a quantile spline (smoothing spline or B-spline) estimator for the data-driven smoothing parameters is still open and it is warranted as a future research topic.

A B-spline approach can be formulated as follows. It is well known that a B-spline approach depends on the degree of smoothness of the true quantile function which determines how well the quantile function can be approximated. Therefore, it is commonly assumed that the quantile function with a certain degree of smoothness r defined as follows. To this end, define a functional

space $\mathcal{Q}_r$ to be the collection of all functions on a domain, say $[0,\ 1]$ for which the m-th order derivative satisfies the Hölder condition of order of γ with $r = m + \gamma$. That is, for each $h \in \mathcal{Q}_r$, $|h^{(m)}(s) - h^{(m)}(t)| \leq W_0|s - t|^\gamma$ for any $0 \leq s,\ t \leq 1$ and a positive finite constant W_0.

Here we first assume that the quantile regression function $q_\tau(x)$ is from $\mathcal{Q}_r$ and then, we can define B-splines of order $m + 1$ used to approximate the quantile function $q_\tau(\cdot)$. We consider a sequence of positive integers $\{k_n\}$, $n \geq 1$, (the number of knots) and an extended partition of $[0,\ 1]$ by k_n knots with equal or unequal length. Then, we can define the associated B-spline basis functions by $\{B_j(x)\}$, $1 \leq j \leq k_n + m + 1$; see Schumaker (1981) for details. The proposed B-spline estimator of $q_\tau(x)$ is given by

$$\widehat{q}_\tau(x) = \sum_{j=1}^{k_n+m+1} \widehat{\theta}_j\, B_j(x),$$

where $\widehat{\theta}_j$ solves the minimization problem

$$\sum_{t=1}^{n} \rho_\tau \left(Y_t - \sum_{j=1}^{k_n+m+1} \theta_j B_j(X_t) \right).$$

Clearly, when the B-spline basis is given, computations can be easily carried using standard quantile regression algorithms as in (5.17). As for selecting the order and knots for the splines, the Schwarz-type information criterion is commonly suggested in the B-spline literature; see He and Shi (1994) and Kim (2007).

5.2.3 Smoothing Parameter Selection

It is well known that the smoothing tuning parameter η ($\eta = h$ for kernel smoothing and $\eta = \lambda$ for smoothing spline) plays an essential role in the trade-off between reducing bias and variance. To the best of our knowledge, there has been very limited literature about selecting η in the context of estimating the quantile regression even though there is a rich amount of literature on this issue in the mean regression setting; see, for example, Cai, Fan and Yao (2000) and Cai and Tiwari (2000). Indeed, Yu and Jones (1998) or Yu and Lu (2004) proposed a simple and convenient method for the nonparametric quantile estimation. Their approach assumes that the second derivatives of the quantile function are parallel. However, this assumption might not be valid for many applications due to (nonlinear) heteroscedasticity. Further, the mean regression approach can not directly estimate the variance function. To attenuate these problems, Cai and Xu (2008) proposed a method of selecting bandwidth for the foregoing estimation procedure, based on the nonparametric version of the Akaike information criterion, which can attend to the structure of time series data and the over-fitting or under-fitting tendency. The basic idea is motivated by its analogue of Cai and Tiwari (2000) for nonlinear mean regression for time series models and we briefly describe it below.

By recalling the classical AIC for linear models under the likelihood setting; that is the negative of twice of the maximized log likelihood plus twice of the number of estimated parameters, Cai and Xu (2008) proposed the following nonparametric version of the bias-corrected AIC; see Hurvich, Simonoff and Tsai (1998) and Cai and Tiwari (2000) for nonparametric regression models, to select η by minimizing

$$AIC(\eta) = \log\left\{\widehat{\sigma}_\eta^2\right\} + 2\,(p_\eta + 1)/[n - (p_\eta + 2)], \tag{5.23}$$

where $\widehat{\sigma}_\eta^2 = n^{-1}\sum_{t=1}^n \rho_\tau\left(Y_t - \widehat{q}_\tau(X_t)\right)$ and p_η is the nonparametric version of degrees of freedom, called the effective number of parameters. This criterion may be interpreted as the AIC for the local quantile smoothing problem and seems to perform well in some limited applications. Note that similar to (5.23), Koenker, Ng and Portnoy (1994) considered the SIC with the second term on the right-hand side of (5.23) replayed by $2\,n^{-1}\,p_\lambda\,\log n$, where p_λ is the number of "active knots" for the smoothing spline quantile setting.

For different smoothing techniques, the choice of p_η might be different. For example, see Koenker, Ng and Portnoy (1994) on how to choose $p_\eta = p_\lambda$ in quantile smoothing splines setting and Cai and Xu (2008) for how to determine $p_\eta = p_h$ under kernel smoothing framework.

5.2.4 Dimension Reduction Modeling

As mentioned earlier, a purely nonparametric quantile regression model may suffer from the so-called "curse of dimensionality" problem. To overcome this difficulty, some dimension reduction modelling methods have been proposed in the literature such as additive and varying-coefficient models, discussed next.

A. Additive Models

An additive quantile regression model takes a form as

$$q_\tau(x) = \delta + \sum_{j=1}^d q_{\tau,j}(x_j), \tag{5.24}$$

which was studied by De Gooijer and Zerom (2003), Yu and Lu (2004), and Horowitz and Lee (2005). For ease of notation, assume that $d = 2$ in what follows. De Gooijer and Zerom (2003) used a two-stage approach to estimate each component in (5.24) as follows. First, estimate the d-dimensional quantile regression surface $g_\tau(x)$ using (5.8) to obtain $\widehat{q}_{\tau,wnw}(x)$ and then use the projection method of Cai and Masry (2000) as

$$\widehat{q}_{\tau,1}(x_1) = \frac{1}{n}\sum_{t=1}^n \widehat{q}_{\tau,wnw}(x_1, X_{t2})W(x_1, X_{t2}),$$

where $W(\cdot)$ is a weighting function, which can be chosen based on minimizing the asymptotic variance as in Cai and Fan (2000) to achieve the optimality or to screen out outliers. Similarly, one can estimate $\widehat{q}_{\tau,2}(x_2)$. De Gooijer and Zerom (2003) also presented the asymptotic normality of the proposed estimator.

Later, Yu and Lu (2004) proposed using a backfitting algorithm equipped with a local linear fitting as follows.

1. Step (1), initial estimation. Set

$$\widehat{\delta} = \text{argmin}_{\delta} \sum_{t=1}^{n} \rho_{\tau}(Y_t - \delta);$$

and, for $j = 1$ and 2,

$$(\widehat{a}_j, \widehat{b}_j) = \text{argmin}_{a,b} \sum_{t=1}^{n} \rho_{\tau}\left(Y_t - \widehat{\delta} - a - b(X_{tj} - x_j)\right) K_{h_j}(X_{tj} - x_j).$$

Then, set $q^{(0)}_{\tau,j}(x_j) = \widehat{a}_j$, and take $q^{*(0)}_{\tau,j}(x_j)$ as $q^{(0)}_{\tau,j}(x_j)$ minus the τ-th sample quantile of $\{q^{(0)}_{\tau,j}(X_{tj})\}_{t=1}^{n}$.

2. Step (2), iteration. Set

$$\widehat{\delta}^{(l)} = \text{argmin}_{\delta} \sum_{t=1}^{n} \rho_{\tau}\left(Y_t - q^{*(l-1)}_{\tau,1}(X_{t1}) - q^{*(l-1)}_{\tau,2}(X_{t2}) - \delta\right);$$

and for $j = 1$ and 2 and $m = 3 - j$,

$$(\widehat{a}_j, \widehat{b}_j) = \text{argmin}_{a,b} \sum_{t=1}^{n} \rho_{\tau}\left(Y_t - \widehat{\delta}^{(l)} - q^{*(l-1)}_{\tau,m}(X_{tm}) - a - b(X_{tj} - x_j)\right) K_{h_j}(X_{tj} - x_j);$$

then take $q^{(l)}_{\tau,j}(x_j) = \widehat{a}_j$, and take $q^{*(l)}_{\tau,j}(x_j)$ as $q^{(l)}_{\tau,j}(x_j)$ minus the τ-th sample quantile of $\{q^{(l)}_{\tau,j}(X_{tj})\}_{t=1}^{n}$.

3. Step (3), keep cycling step (2) for $l = 1, 2, 3, \cdots$ until the value of $q^{*(l)}_{\tau} = (\widehat{\delta}^{(l)}, q^{*(l)}_{\tau,1}, q^{*(l)}_{\tau,2})$ has converged. Next, for $j = 1$ and 2, let $(\widehat{a}_j, \widehat{b}_j) = (q^{*(l)}_{\tau,j}(x_j), \widehat{b}_j)$. Then, $(\widehat{a}_j, \widehat{b}_j)$ gives the estimators of $q_{\tau,j}(x_j)$ and $q'_{\tau,j}(x_j)$, respectively.

Further, Yu and Lu (2004) investigated the large sample behavior of the proposed backfitting estimator.

Recently, Horowitz and Lee (2005) used a two-stage approach which is different from that in De Gooijer and Zerom (2003). At the first stage, use a series approximation to each component as $q_{\tau,j}(x_j) \approx \sum_{l=0}^{k_j} \theta_{lj}\phi_{jl}(x_j)$, where $\{\phi_{jl}(\cdot)\}$ is a basis function, and then estimate θ_{lj} by

$$\text{argmin}_{\delta,\theta} \sum_{t=1}^{n} \rho_{\tau}\left(Y_t - \delta - \sum_{j=1}^{2}\sum_{l=0}^{k_j} \theta_{lj}\phi_{jl}(x_j)\right),$$

denoted by $\widehat{\theta}_{lj}$, to obtain

$$\widehat{q}^{(0)}_{\tau,j}(x_j) = \sum_{l=0}^{k_j} \widehat{\theta}_{lj}\phi_{jl}(x_j).$$

At the second stage, estimate $q_{\tau,j}(x_j)$ by first finding

$$(\widehat{a}_j, \widehat{b}_j) = \text{argmin}_{a,b} \sum_{t=1}^{n} \rho_\tau \left(Y_t - \widehat{\delta} - \widehat{q}^{(0)}_{\tau,m}(X_{tm}) - a - b(X_{tj} - x_j) \right) K_{h_j}(X_{tj} - x_j);$$

and then taking $\widehat{q}_{\tau,j}(x_j) = \widehat{a}_j$. Also, Horowitz and Lee (2005) derived the asymptotic properties for the proposed two-stage estimator.

B. Varying-Coefficient Models

A varying-coefficient quantile regression model takes a form as

$$q_\tau(u, x) = \sum_{j=1}^{d} a_{\tau,j}(u)\, x_j = a_\tau(u)^T\, x, \tag{5.25}$$

which was studied by Honda (2004) for iid data, Cai and Xu (2008) for dynamic time series observations, and Kim (2007) for time-varying coefficients (u is time) for iid samples. For easy exposition, we assume that u is univariate below.

To estimate $\{a_k(\cdot)\}$ using the local polynomial method based on $\{(U_t,\, X_t,\, Y_t)\}_{t=1}^n$, assume that the coefficient functions $\{a(\cdot)\}$ have the $(m+1)$th derivative ($m \geq 1$), so that for any given gird point $u \in \Re$, $a_k(\cdot)$ can be approximated by a polynomial function in a neighborhood of the given grid point u as $a(U_t) \approx \sum_{j=0}^m \beta_j\, (U_t - u)^j$, where $\beta_j = a^{(j)}(u)/j!$ and $a^{(j)}(u)$ is the jth derivative of $a(u)$, so that $q_\tau(U_t,\, X_t) \approx \sum_{j=0}^m X_t^T\, \beta_j\, (U_t - u)^j$. Then, the locally weighted loss function is

$$\sum_{t=1}^{n} \rho_\tau \left(Y_t - \sum_{j=0}^{m} X_t^T\, \beta_j\, (U_t - u)^j \right) K_h(U_t - u). \tag{5.26}$$

Solving the minimization problem in (5.26) gives $\widehat{a}(u) = \widehat{\beta}_0$, the local polynomial estimate of $a(u)$, and $\widehat{a}^{(j)}(u) = j\,!\,\widehat{\beta}_j$ $(j \geq 1)$, the local polynomial estimate of the jth derivative $a^{(j)}(u)$. By moving u along with the real line, the estimate of the entire curve $\widehat{a}(u)$ is obtained.

Cai and Xu (2008) derived the asymptotic properties for $\widehat{a}(u)$. Under some regularity conditions, we have the following asymptotic normality for m odd,

$$\sqrt{nh} \left[\widehat{a}(u) - a(u) - \frac{h^{m+1}}{(m+1)!} a^{(m+1)}(u) \mu_{m+1}(K) + o_p(h^{m+1}) \right] \xrightarrow{d} N\left(0, \Sigma_a(u)\right),$$

where $\Sigma_a(u) = \tau(1-\tau)\Sigma(u)$, $\Sigma(u) = [\Omega^*(u)]^{-1}\, \Omega(u)\, [\Omega^*(u)]^{-1} / f_u(u)$, $\Omega(u) = E[X_t X_t^T | U_t = u]$, $\Omega^*(u) = E[X_t X_t^T\, f_{y|u,x}(q_\tau(u, X_t))\,|\, U_t = u]$, $f_u(\cdot)$ is the marginal density of U_t, and $f_{y|u,x}(y)$ is the conditional density of Y_t given U_t and X_t. Also, Cai and Xu (2008) proposed an ad hoc bandwidth selection method which is similar to that described in Section 5.2.3.

Finally, Kim (2007) considered the time-varying coefficient quantile regression model as

$$q_\tau(t,x) = \sum_{j=1}^{d} a_{\tau,j}(t)\, x_j = a_\tau(t)^T\, x, \tag{5.27}$$

and used a B-spline technique to estimate $a_\tau(t)$. Note that model (5.27) might be potentially useful to see whether the quantile regression changes over time and in a case with a practical interest is, for example, the analysis of the reference growth data by Cole (1994), Wei, Pere, Koenker and He (2006), and Wei and He (2006) for longitudinal data, and Kim (2007) for iid samples. Finally, it is worth to point out that model (5.27) might be very useful for a nonparametric testing for testing structural changes in regression quantiles as in Qu (2008).

Cai and Xu (2008) used model (5.25) and its modeling approaches to explore the possible nonlinearity feature, heteroscedasticity, and predictability of the exchange rate series of the Japanese Yen in terms of the U.S. dollar. Their empirical findings are that the quantile has a complex structure and that both heteroscedasticity and nonlinearity exist. This implies that the GARCH effects occur in the exchange rate time series. Finally, they considered the one-step ahead post-sample forecasting for the last 25 observations and constructed the 95% nonparametric prediction interval $(\widehat{q}_{.025}(\cdot), \widehat{q}_{.975}(\cdot))$ based on the past two lags. It turns out that 24 of 25 predictive intervals contain the corresponding true values. This means that under the dynamic smooth coefficient quantile regression model assumption, the prediction intervals based on the proposed method work reasonably well.

6 Conclusion

In this paper we survey some recent developments in nonparametric econometrics, including (i) nonparametric estimation and testing of regression functions with mixed discrete and continuous covariates; (ii) nonparametric estimation/testing with nonstationary data; (iii) nonparametric instrumental variable estimations; and (iv) nonparametric estimation of quantile regression models.

In the paper by Cai and Hong (2009), they gave a survey on the recent developments of nonparametric estimation and testing of financial econometric models. Due to space limitation we omit some of the important areas such as nonparametric/semiparametric with limited dependent variable models and nonparametric/semiparametric panel data models. Another promising line of research is to impose less restrictions on econometric models and hence parameters may not be point identified but are set identified. Readers interested in these areas of research should consult with the works by Manski (2003), Imbens and Manski (2004), Honore and Tamer (2006) and the references therein.

References

Ai, C. and X. Chen (2003). Efficient estimation of models with conditional moment restrictions conditioning unknown functions. *Econometrica*, **71**, 1795-1843.

Aitchison, J. and Aitken, C.G.G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, **63**, 413-420.

Bachmeier, L., S. Leelahanon and Q. Li (2006). Money growth and inflation in the United States. *Macroeconomic Dynamics*, **11**, 113-127.

Blundell, R. and J. Powell (2003). Endogeneity in nonparametric and semiparametric regression models. In *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, Vol. II (M. Dewatripont, L.P. Hansen and S.J. Turnovsky, eds.), Cambridge: Cambridge University Press.

Cai, Z. (2002a). Two-step likelihood estimation procedure for varying-coefficient models. *Journal of Multivariate Analysis*, **81**, 189-209.

Cai, Z. (2002b). Regression quantile for time series. *Econometric Theory*, **18**, 169-192.

Cai, Z. (2002c). A two-stage approach to additive time series models. *Statistica Neerlandica*, **56**, 415-433.

Cai, Z., M. Das, H. Xiong and X. Wu (2006). Functional coefficient instrumental variables models. *Journal of Econometrics*, **133**, 207-241.

Cai, Z. and J. Fan (2000). Average regression surface for dependent data. *Journal of Multivariate Analysis*, **75**, 112-142.

Cai, Z., J. Fan and R. Li (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, **95**, 888-902.

Cai, Z., J. Fan and Q. Yao (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, **95**, 941-956.

Cai, Z. and Y. Hong (2009). Some recent developments in nonparametric finance. Forthcoming in *Advances in Econometrics*.

Cai, Z., Q. Li and J. Park (2009). Functional-coefficient models for nonstationary time series Data. *Journal of Econometrics*, **148**, 101-113.

Cai, Z. and E. Masry (2000). Nonparametric estimation in nonlinear ARX time series models: Projection and linear fitting. *Econometric Theory*, **16**, 465-501.

Cai, Z. and E. Ould-Said (2003). Local M-estimator for nonparametric time series. *Statistics and Probability Letters*, **65**, 433-449.

Cai, Z. and R.C. Tiwari (2000). Application of a local linear autoregressive model to BOD time series. *Environmetrics*, **11**, 341-350.

Cai, Z. and X. Wang (2008). Nonparametric methods for estimating conditional value-at-risk and expected shortfall. *Journal of Econometrics*, **147**, 120-130 .

Cai, Z. and Y. Wang (2009). Instability of predictability of assets returns. *Working paper*, University of North Carolina at Charlotte.

Cai, Z. and H. Xiong (2006). Partially varying coefficient instrumental variable models. *Working paper*, University of North Carolina at Charlotte.

Cai, Z. and X. Xu (2008). Nonparametric quantile estimations for dynamic smooth coefficient models. *Journal of the American Statistical Association*, **103**, 1595-1608.

Caner, M. and B.E. Hansen (2004). Instrumental variable estimation of a threshold model. *Econometric Theory*, **20**, 813-843.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, **69**, 1127-1160.

Chaudhuri, P. (1991). Nonparametric estimates of regression quantiles and their local Bahadur representation. *The Annals of Statistics*, **19**, 760-777.

Chaudhuri, P., K. Doksum and A. Samarov (1997). On average derivative quantile regression. *The Annals of Statistics*, **25**, 715-744.

Cole, T.J. (1994). Growth charts for both cross-sectional and longitudinal data. *Statistics in Medicine*, **13**, 2477-2492.

Daroles, S., J.-P. Florens and E. Renault (2002). Nonparametric instrumental regression. Working Paper, GREMAQ, University of Social Science, Toulouse.

Das, M. (2003). Identification and sequential estimation of nonparametric panel models with insufficient exclusion restrictions. *Journal of Econometrics*, **114**, 297-328.

Das, M. (2005). Instrumental variables estimators for nonparametric models with discrete endogenous regressors. *Journal of Econometrics*, **124**, 335-361.

Das, M., W. Newey and F. Vella (2003), Nonparametric estimation of sample selection models. *The Review of Economic Studies*, **70**, 33-58.

De Gooijer, J. and D. Zerom (2003). On additive conditional quantiles with high dimensional covariates. *Journal of the American Statistical Association*, **98**, 135-146.

Engle, R.F. and S. Manganelli (2004). CAViaR: Conditional autoregressive value at risk by regression quantile. *Journal of Business and Economics Statistics*, **22**, 367-381.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modeling and Its Applications*. Chapman and Hall, London.

Fan, J., T.-C. Hu and Y.K. Troung (1994). Robust nonparametric function estimation. *Scandinavian Journal of Statistics*, **21**, 433-446.

Fan, J. and T. Huang (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, **11**, 1031-1057.

Fan, Y. and Q. Li (1996). Consistent model specification tests: omitted variables and semiparametric functional forms. *Econometrica*, **64**, 865-890.

Gao, J., M. King, Z. Lu and D. Tjøstheim (2008). Nonparametric specification testing for nonlinear time series with nonstationarity. Manuscript.

Gu, J. and P. Hernandez-Verme (2009). An empirical evaluation of the presence of credit rationing in the U.S. credit markets. Manuscript.

Hall, P. and J.L. Horowitz (2005). Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, **33**, 2904-2929.

Hall, P., Q. Li and J. Racine. (2007). Nonparametric estimation of regression functions in the presence of irrelevant regressors. *Review of Economic and Statistics*, **89**, 784-789.

Hall, P., J. Racine and Q. Li (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, **99**, 1015-1026.

He, X. and P. Ng (1999). Quantile splines with several covariates. *Journal of Statistical Planning and Inference*, **75**, 343-352.

He, X., P. Ng and S. Portnoy (1998). Bivariate quantile smoothing splines. *Journal of the Royal Statistical Society, Series B*, **60**, 537-550.

He, X. and S. Portnoy (2000). Some asymptotic results on bivariate quantile splines. *Journal of Statistical Planning and Inference*, **91**, 341-349.

He, X. and P. Shi (1994). Convergence rate of B-spline estimators of nonparametric conditional quantile functions. *Journal of Nonparametric Statistics*, **3**, 299-308.

Honda, T. (2000). Nonparametric estimation of a conditional quantile for α-mixing processes. *Annals of the Institute of Statistical Mathematics*, **52**, 459-470.

Honda, T. (2004). Quantile regression in varying coefficient models. *Journal of Statistical Planning and Inferences*, **121**, 113-125.

Honore, B. and E. Tamer (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica*, **74**, 611-630.

Horowitz, J.L. (2007). Asymptotic normality of a nonparametric instrument variables estimator. *International Economic Review*, **48**, 1329-1349.

Horowitz, J.L. and S. Lee (2005). Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association*, **100**, 1238-1249.

Hsiao, C., Q. Li and J. Racine (2007). Consistent model specification tests with mixed discrete and continuous variables. *Journal of Econometrics*, **140**, 802-826.

Hurvich, C.M., J.S. Simonoff and C.-L. Tsai (1998). Smoothing parameter selection in non-parametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B*, **60**, 271-293.

Imbens, G. and C. Manski (2004). Confidence intervals for partially identified parameters. *Econometrica*, **72**, 1845-1857.

Juhl, T. (2005). Functional coefficient models under unit root behavior. *Econometrics Journal*, **8**, 197-213.

Karlsen, H.A., T. Myklebust and D. Tjöstheim (2007). Nonparametric estimation in a nonlinear cointegration type model. *The Annals of Statistics*, **35**, 252-299.

Khindanova, I.N. and S.T. Rachev (2000). Value at risk: Recent advances. *Handbook on Analytic-Computational Methods in Applied Mathematics*, CRC Press LLC.

Kim, M.-O. (2007). Quantile regression with varying coefficients. *The Annals of Statistics*, **35**, 92-108.

Koenker, R. (2004). *Quantreg: An R package for quantile regression and related methods* **http://cran.r-project.org**.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monograph Series, Cambridge University Press.

Koenker, R. and G.W. Bassett (1978). Regression quantiles. *Econometrica*, **46**, 33-50.

Koenker, R. and G.W. Bassett (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, **50**, 43-61.

Koenker, R., P. Ng and S. Portnoy (1994). Quantile smoothing splines. *Biometrika*, **81**, 673-680.

Koenker, R., S. Portnoy and P. Ng (1992). Nonparametric estimation of conditional quantile functions. In L_1-*Statistical Analysis and Related Methods* (Y. Dodge, ed.). Amsterdam: Elsevier, 217-229.

Li, Q., C. Hsiao and J. Zinn (2003). Consistent model specification tests for semiparametric/nonparametric models based on series estimation methods. *Journal of Econometrics*, **112**, 295-325.

Li, Q., C. Huang, D. Li and T. Fu (2002). Semiparametric smooth coefficient models. *Journal of Business and Economic Statistics*, **20**, 412-422.

Li, Q., D. Ouyang and J. Racine (2009). Nonparametric regression with weakly dependent data: The discrete and continuous regressor case. Forthcoming in *Journal of Nonparametric Statistics*.

Li, Q. and J. Racine (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University, Princeton and Oxford.

Li, Q. and J. Racine (2009). Smooth varying-coefficient estimation and inference for qualitative and quantitative Data. Under revision for *Econometric Theory*.

Liang, Z. and Q. Li (2009). Functional Coefficient Regression Model with Time Trend. Manuscript.

Manski, C. (2003). *Partial Identification of Probability Distributions*. Springer-Verlag, New York.

Masry, E. and D. Tjøstheim (1997). Additive nonlinear ARX time series and projection estimates. *Econometric Theory*, **13**, 214-252.

Newey, W.K. and J.L. Powell (1988). Instrumental variables estimation for nonparametric models. Manuscript.

Newey, W.K. and J.L. Powell (2003). Nonparametric instrumental variables estimation. *Econometrica*, **71**, 1565-1578.

Newey, W.K., J.L. Powell and F. Vella (1999). Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, **67**, 565-603.

Ouyang, D., Q. Li and J. Racine (2009). Nonparametric estimation of regression functions with discrete regressors. *Econometric Theory*, **25**, 1-42.

Pakes, A. and S. Olley (1995). A limit theorem for a smooth class of semiparametric estimators. *Journal of Econometrics*, **65**, 295-332.

Park, J.Y. and S.B. Hahn (1999). Cointegrating regressions with time varying coefficients. *Econometric Theory*, **15**, 664-703.

Park, S. (2003). Semiparametric instrumental variables estimation. *Journal of Econometrics*, **112**, 381-399.

Phillips, P.C.B. and J. Park (1998). Nonstationary density estimation and kernel autoregression. Manuscript.

Portnoy, S. (1997). Local asymptotics for quantile smoothing splines. *The Annals of Statistics*, **25**, 414-434.

Racine, J., J. Hart and Q. Li (2006). Testing the significance of categorical variables. *Econometric Reviews*, **25**, 523-544.

Racine, J. and Q. Li, (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, **119**, 99-130.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *Journal of the American Statistical Association*, **92**, 1057-1062.

Ruppert, D., S.J. Sheather and M.P. Wand. (1995). An effective bandwidth selection for local least squares regression. *Journal of the American Statistical Association*, **90**, 1257-1270.

Qu, Z. (2008). Testing for structural change in regression quantiles. Forthcoming in *Journal of Econometrics*.

Schultz, T.P. (1997). *Human Capital, Schooling and Health*. IUSSP, XXIII, General Population Conference, Yale University.

Schumaker, L.L. (1981). *Spline Functions: Base Theory*. Wiley, New York.

Speckman, P. (1988). Kernel smoothing partial linear models. *The Journal of Royal Statistical Society, Series B*, **50**, 413-426.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B*, **36**, 111-147.

Su, L., Y. Chen and A. Ullah (2009). Functional coefficient estimation with both categorical and continuous data. Forthcoming in *Advances in Econometrics*.

Su, L. and A. Ullah (2008). Local polynomial estimation of nonparametric simultaneous equations models. *Journal of Econometrics*, **144**, 193-218.

Sun, Y., Z. Cai and Q. Li (2008). Consistent nonparametric test on parametric smooth coefficient model with nonstationary data. Manuscript.

Sun, X., C. Hsiao and Q. Li (2008). Volatility Spillover Effect: A Semiparametric Analysis of Non-Cointegrated Processes. Manuscript.

Sun, Y. and Q. Li (2009). Data-Driven Method Selecting Smoothing Parameters in Semiparametric Models with Integrated Time Series Data. Manuscript.

Sun, Y. and Q. Li (2009). Cointegration test on semiparametric smooth coefficient models. Manuscript.

Wang, Q., Phillips, P.C.B. (2006). Asymptotic theory for local time density estimation and nonparametric cointegrating regression. Forthcoming in *Econometric Theory*.

Wang, Q., Phillips, P.C.B. (2008). Structural nonparametric cointegrating regression. Manuscript.

Wei, Y. and X. He (2006). Conditional growth charts (with discussion). *The Annals of Statistics*, **34**, 2069-2097.

Wei, Y., A. Pere, R. Koenker and X. He (2006). Quantile regression methods for reference growth charts. *Statistics in Medicine*, **25**, 1369-1382.

Xiao, Z. (2009). Functional coefficient co-integration models. Forthcoming in *Journal of Econometrics*.

Yu, K. and M.C. Jones (1998). Local linear quantile regression. *Journal of the American Statistical Association*, **93**, 228-237.

Yu, K. and Z. Lu (2004). Local linear additive quantile regression. *Scandinavian Journal of Statistics*, **31**, 333-346.

Zheng, J.X. (1996). A consistent test for functional form via nonparametric estimation techniques. *Journal of Econometrics*, **75**, 263-289.